

TARTU ÜLIKOOL
BIOLOOGIA-GEOGRAAFIATEADUSKOND
MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT
BIOINFORMAATIKA ÕPPETOOL

Age Tats

**Nukleotiidide, koodonite ja aminohapete sagedused
kõrge ekspressioonitasemega geenides**

Magistritöö

Juhendajad prof. Maido Remm, PhD
Tanel Tenson, PhD

TARTU 2006

SISUKORD

KASUTATUD LÜHENDID	4
SISSEJUHATUS	7
KIRJANDUSE ÜLEVAADE	8
1. Translatsiooni initsiatsiooni mehhanismid.....	8
1.1 Bakterid	8
1.2 Eukarüootid	8
1.3 Arhead	9
2. Valgu N-terminaalsed signaalid	10
2.1 Lokalisatsioonisignaalid.....	10
2.2 Amino-terminaalse metioniini töötlemine	10
2.3 Degradatsioonisignaalid.....	11
3. Translatsiooni initsiatsiooni mõjutavad järjestuse motiivid.....	13
3.1 Bakterid	13
3.1.1 Startkoodonist ülesvoolu asuvad motiivid	14
3.1.2 Startkoodonist allavoolu asuvad motiivid.....	14
3.2 Eukarüootid	15
3.2.1 Startkoodonit ümbritseva järjestuse kontekst	16
3.3 Arhead	16
4. Koodonkasutus ja geenide ekspresioonitase.....	17
4.1 Koodoneelistuse mõõtmise meetodid	17
4.1.1 CAI.....	17
4.1.2 Koodonkasutuse erinevuste indeks	19
4.2 Koodonkasutuse ja peptidüül-tRNA <i>drop-off</i> i seos.....	20
PRAKTIINE OSA	21
TÖÖ EESMÄRK	21
ANDMED JA MEETODID	21
1. Andmed	21
1.1 Kõrge ekspresioonitasemega geenid üherakulistes organismides	22
1.2 Kõrge ekspresioonitasemega geenid inimeses	23
2. Meetodid	23
2.1 A nukleotiidi sageduse ja CAI seose analüüs	23
2.2 Nukleotiidide eelistuse analüüs.....	24
2.3 Koodonite eelistuse analüüs.....	24
2.4 Aminohapete eelistuse analüüs	24
3. Statistiline olulisus	25
TULEMUSED JA ARUTELU	26
1. Adenosiini sagedus <i>E. coli</i> avatud lugemisraamide alguses	26
2. Nukleotiidide kasutus erinevate üherakuliste organismide ORF-ide alguses	27
3. Koodonite eelistused üherakulistes organismides.....	31
4. Aminohapete eelistused üherakulistes organismides.....	35
5. Järjestuse analüüs inimese kõrge ekspresioonitasemega ORF-ide alguses.....	38
6. Alandmestiku valiku kontroll.....	40

7. Ortoloogidel põhineva kõrgelt ekspresseeruvate geenide leidmise meetodi kontroll..	41
KOKKUVÕTE.....	43
SUMMARY	45
TÄNUAVALDUSED	47
VIITED	48
LISAD	57

KASUTATUD LÜHENDID

A	adenosiin
Ala	alaniin
Arg	arginiin
A-sait	aminoatsüülsait
Asn	asparagiin
Asp	aspartaat
C	tsütidiin
CAI	koodoni adaptatsiooni indeks (<i>codon adaptation index</i>)
CBI	koodoneelistuse indeks (<i>codon bias index</i>)
CCDS	konsensus kodeeriv järjestus (<i>consensus coding sequence</i>)
CDS	kodeeriv järjestus (<i>coding sequence</i>)
C-terminaalne	karboksüülterminaalne
Cys	tsüsteiin
DR	allavoolu asuv regioon (<i>downstream region</i>)
EF	elongatsioonifaktor
ENC	efektiivne koodonite arv (<i>effective number of codons</i>)
FET	Fisher'i täpne test (<i>Fisher's exact test</i>)
fMet	formüül-metioniin
fMet-tRNA ^{fMet}	N-formüül-metionüül-tRNA
Fop	optimaalsete koodonite sagedus (<i>frequency of optimal codons</i>)
G	guanosiin
Gln	glutamiin
Glu	glutamaat

Gly	glütsiin
H ₀	nullhüpotees
IF	initsiatsioonifaktor
Ile	isoleutsiin
IRES	sisemine ribosoomi sisenemiskoht (<i>internal ribosomal entry site</i>)
KEG	kõrge ekspressioonitasemega geenid (geenid, millelt transleeritakse valku kõrge efektiivsusega)
Leu	leutsiin
Lys	lüsiin
m ⁷ G	7-metüülguanosiin
MAP	metioniini aminopeptidaas
Mbp	miljon aluspaari (<i>mega basepairs</i>)
Met	metioniin
mRNA	informatsiooni (<i>messenger</i>) RNA
N-formüül	amino-formüül
nt	nukleotiid
N-terminaalne	aminoterminalne
ORF	avatud lugemisraam (<i>open reading frame</i>)
Phe	fenüülalaniin
Pro	proliin
P-sait	peptidüülsait
p-väärtus	tõenäosusväärtus (<i>probability value</i>)
RNA	ribonukleiinhape
rRNA	ribosomaalne RNA
SAGE	geeniekspressiooni seeriaanalüüs (<i>serial analysis of gene expression</i>)
SD	Shine-Dalgarno

Ser	seriin
Thr	treoniin
tRNA	transport RNA
Trp	trüptofaan
Tyr	türosiin
U	uridiin
uORF	ülesvoolu asuv avatud lugemisraam (<i>upstream ORF</i>)
Val	valiin

SISSEJUHATUS

Geeniekspressiooni viimase etapi, translatsiooni, põhilised mehhanismid ja komponendid on evolutsiooniliselt konserveerunud. Siiski esinevad ulatuslikud erinevused bakterite ja eukarüootide translatsiooni initsiatsioonis – peamises translatsiooni efektiivsust määravas etapis. Bakteriaalne translatsiooni initsiatsioon põhineb otsesel translatsiooni initsiatsiooniregiooni äratundmisel ja sellega seostumisel ribosoomi poolt. Eukarüootides toimib aga keerukas ja mitmeid abifaktoreid vajav mehhanism, kus ribosoomi väiksem alaühik libiseb mööda mRNA-d kuni sobiva startkoodonini. Arheate translatsiooni puhul esineb nii eukarüootsele kui ka bakteriaalsele translatsiooni initsiatsioonile iseloomulikke omadusi.

Translatsiooni initsiatsiooni efektiivsuse määrab initsiatsiooniregiooni edukas äratundmine. Selle äratundmise eest vastutavad mitmed mRNA järjestuse motiivid, mis asuvad initsiatsioonikoodonist nii üles- kui allavoolu. Nii on näiteks tänaseks teada, et startkoodonile vahetult järgnev järjestuse kontekst on mittejuhuslik ning geenide ekspressioonitaseme ja koodonkasutuse vahel esineb tugev korrelatsioon. Selliseid spetsiifilisi järjestuse omadusi on uuritud paljudes töödes, kuid tavaliselt hõlmavad tehtud analüüsid organismi kõiki geene. Geenide ekspressioonitasemeid mitte arvestades võivad märkamata jääda spetsiifilised, ainult kõrge ekspressioonitasemega geenide gruppide omased järjestuse motiivid. Eelnevat arvesse võttes on käesolevas töös analüüsitud startkoodonist allavoolu asuvaid mRNA regioone üheksas bakteris, kolmes arheas, 3 üherakulises eukarüootis ja inimeses, võrreldes seejuures kõrge ekspressioonitasemega geenide gruppi organismi kõigi geenide grupiga. Nukleotiidide, koodonite ja aminohapete sageduste detailsete analüüside eesmärgiks on välja selgitada, kas kõrge ekspressioonitasemega geenid omavad vahetult startkoodonile järgnevates järjestustes iseloomulikku mustrit.

KIRJANDUSE ÜLEVAADE

1. Translatsiooni initsiatsiooni mehhanismid

1.1 Bakterid

Translatsiooni initsiatsioon bakterites algab ribosoomi 30S subühiku 16S rRNA 3' otsa seondumisega mRNA-l asuva SD järjestusega (Shine ja Dalgarno, 1974). Järgnevalt siseneb fMet-tRNA^{fMet} ribosoomi P-saiti. Mõlemat protsessi soodustavad kolm initsiatsioonifaktorit: IF1, IF2, IF3. Moodustunud 30S initsiatsioonikompleks koosneb 30S subühikust, fMet-tRNA^{fMet}-st ja kolmest initsiatsioonifaktorist (Gualerzi ja Pon, 1990). 30S initsiatsioonikompleksi moodustumine on valgusünteesi taset määrav etapp ning seetõttu peamine translatsiooni efektiivsuse määraja (de Smit ja van Duin, 1994). Elongatsioonifaktori EF-Tu abil toimetatakse ribosoomi A-saiti mRNA teise koodoniga komplementaarne aminoatsüül-tRNA.

1.2 Eukarüooidid

Eukarüootne translatsiooni initsiatsioon võib toimuda *cap*-sõltuvalt või sõltumatult, vastavalt sellele, kuidas ribosoomi 40S subühik seotakse mRNA-ga.

Cap-sõltuva translatsiooni initsiatsiooni puhul seondub ribosoomi väike subühik (40S) *cap* struktuurile, mis asub mRNA 5' otsas, ning skaneerib mRNA-d kuni leiab esimese sobivas järjestuse kontekstis asuva startkoodoni (Kozak, 1995), mis tuntakse ära aluspaardumise kaudu Met-tRNAⁱ-ga (Cigan jt., 1988).

Esimese AUG reeglist möödapääsemiseks esineb eukarüootides kolm erinevat *cap*-sõltumatut mehhanismi, mis lubavad translatsiooni alustada teiselt või harvemini kolmandalt mRNA 5' otsa lähedal paiknevalt AUG koodonilt:

- 1) nn. „*leaky scanning*“ mehhanism – ribosoomi väike subühik võib esimesest AUG koodonist mööduda, kui see asub liiga lähedal 5' otsale ja seetõttu ei suudeta teda

efektiivselt ära tunda, samuti võib antud mehhanismi põhjuseks olla initsiatsioonikoodoni halb kontekst (Kozak, 1991; Slusher jt., 1991).

- 2) reinitsiatsioon – kui 80S ribosoom on transleerinud esimese väikese avatud lugemisraami (uORF-i) ja jõudnud terminatsioonikoodonini, võib 40S subühik jääda mRNA-le, jätkata skaneerimist ning reinitseerida AUG koodonist allavoolu (Kozak, 1999).
- 3) ribosomaalne „*shunting*“ – 40S subühik hüppab üle mRNA sekundaarstruktuuride, mis takistaksid skaneerimist (Curran ja Kolakofsky, 1988; Yueh ja Schneider, 1996).

Lisaks arvatakse, et eukarüootne ribosoom võib otse seonduda teatud mRNA-de sisemistele järjestustele (IRES) (Le ja Maizel, 1997; Martinez-Salas, 1999).

1.3 Arhead

Eksperimentaalselt on arheates näidatud kahe erineva initsiatsioonimehhanismi olemasolu – üks liidriga ja teine liidrita mRNA-de jaoks. Polütsistronsete mRNA-de distaalsetel tsistronitel kasutatakse SD-sõltuvat initsiatsiooni ning polütsistronsete tsistronite esimese tsistroni ning monotsistronsete mRNA-de puhul kasutatakse liidrita initsiatsiooni (Benelli jt., 2003). Liidriga ja liidrita transkriptide ning seega ka erinevate translatsioonimehhanismide osakaal on erinevates organismides erinev (Torarinsson jt., 2005). Lisaks on kõigist siiani sekveneeritud arheate genoomidest leitud eukarüootsete initsiatsioonifaktorite homolooge.

Seega on arheate translatsiooni initsiatsioonil nii bakteriaalsele kui ka eukarüootsele translatsiooni initsiatsioonile iseloomulikke omadusi.

2. Valgu N-terminaalsed signaalid

2.1 Lokalisatsioonisignaalid

Valkude N-terminaalses osas paiknevad mitmed erinevad signaalid, mis osalevad olulistes valgu eluiga ja funktsioneerimist mõjutavates mehhanismides. Nn. lokalisatsioonisignaalid, mis võivad varieeruda 15 kuni 50 aminohappejäägini (von Heijne, 1986), suunavad bakterites tekkivad või täielikult valmis valgud tsütosoolist plasmamembraani (Schatz ja Beckwith, 1990). Signaaljärjestused võivad asuda ka valgu sees või C-terminaalses otsas (Kutay jt., 1995). Lokalisatsioonijärjestused ei puuduta otseselt paari esimest N-terminaalset aminohappejääki ning seega jäävad käesolevas töös lähemalt uuritud piirkonnast välja.

2.2 Amino-terminaalse metioniini töötlemine

Bakteriaalse valgusünteesi alguses moodustub polüpeptiidile fMet ots (Kozak, 1983; Meinnel jt., 1993; Schmitt jt., 1996), mis ei ole valgu initsiatsiooniks absoluutselt vajalik ning seda võib vaadelda kui initsiatsiooni optimeerimise strateegiat (Mazel jt., 1994). fMet töödeldakse järgneva elongatsiooni käigus. Kõigepealt eemaldatakse N-formüül osa deformülaasiga, mille tulemusena jääb aminoterminusse metioniin. Märkimisväärse hulga valkude puhul eemaldatakse järgnevalt ka aminoterminalne metioniin (ülevaade (Sherman jt., 1985)). Erinevates töödes on näidatud, et vastav ensüüm – MAP – lõikab metioniini ära juhul kui metioniin esineb selliste aminohapete ees nagu Ala, Gly, Pro, Ser, Thr või Val, ning lõikamist ei toimu, kui metioniinile vahetult järgnev aminohape on Arg, Asn, Asp, Gln, Glu, Ile, Leu, Lys või Met (Tsunasawa jt., 1985; Ben-Bassat jt., 1987; Miller jt., 1987; Moerschell jt., 1990). Tavaliselt eelistatakse lõikusel metioniinile järgnevat jääki, mille külghel on lühike; pika või koguka külghela puhul pole MAP võimeline metioniini eemaldama (Hirel jt., 1989; Dalboge jt., 1990; Schmitt jt., 1996).

Lisaks ei suuda MAP eemaldada ka deformüleerimata metioniini ning seetõttu on alati vajalik eelnev N-formüül-metioniini deformülatsioon (Solbiati jt., 1999).

N-terminaalse metioniini ära lõikamine toimub ka eukarüootides ja arheates. *Saccharomyces cerevisiae*'s ja inimeses esineb kaks MAP-i: MetAP1 ja MetAP2; MetAP1 on vähem efektiivne metioniini eemaldamisel Val-i eest ning MetAP2 omakorda metioniini eemaldamisel Gly-i, Cys-i või Thr-i eest (Chen jt., 2002). Arheate MAP paikneb järjestuse alusel evolutsiooniliselt eukarüootide ja bakterite MAP-ide vahel (Tsunasawa jt., 1997).

2.3 Degradatsioonisignaalid

N-terminaalne reegel (*N-end rule*) seob valgu eluea tema N-terminaalsete aminohappejääkidega. Valkude omadusi, mis tingivad molekuli lagundamise, kutsutakse degradatsioonisignaalideks ehk N-degroniteks. Peamine degradatsioonisignaali komponent on valgu destabiliseeriv N-terminaalne aminohappejääk. Nimelt teatud aminohapete esinemine valgu N-terminaalses otsas suunab valgu degradatsiooni. Destabiliseerivad aminohappejäägid jagatakse primaarseteks, sekundaarseteks ja tertsiaarseteks (tabel 1). Tertsiaarsete destabiliseerivate jääkide aktiivsus vajab juurdepääsu N-terminaalsele amidohüdrolaasile, mille abil tertsiaarsed jäägid muudetakse sekundaarseteks. *E. coli*'s tertsiaarsed destabiliseerivad jäägid puuduvad. Sekundaarsed destabiliseerivad aminohappejäägid muudetakse bakterite puhul Leu/Phe-tRNA-proteiini transferaasi ning eukarüootide puhul Arg-tRNA-proteiini transferaasi abil omakorda primaarseteks destabiliseerivateks aminohappejääkideks. Primaarsete jääkide destabiliseeriv aktiivsus avaldub nende seondumisel ühega kahest valgu E3 seondumiskohast; vastavalt sellele jagatakse primaarsed destabiliseerivad jäägid omakorda tüüp 1 ja tüüp 2 primaarseteks destabiliseerivateks jääkideks. Olemas on ka stabiliseerivad jäägid, millele N-terminaalse raja sihtmärkkomponendid ei seonu piisavalt efektiivselt või ei modifitseeri neid piisavalt efektiivselt (Varshavsky, 1996).

Tabel 1. Destabiliseerivad ja stabiliseerivad aminohappejäägid *E. coli*'s, *S. cerevisiae*'s ja imetajates (küüliku retikulotsüütides ja hiire L-rakkudes) (Varshavsky, 1996). Punane – primaarsed destabiliseerivad tüüp 1, sinine – primaarsed destabiliseerivad tüüp 2.

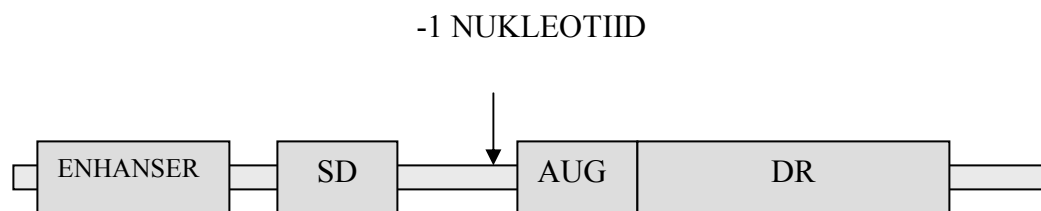
	<i>E.coli</i>	<i>S.cerevisiae</i>	Imetajad
Destabiliseerivad			
Primaarsed	Phe Leu Trp Tyr	Phe Leu Trp Tyr Ile Arg Lys His	Phe Leu Trp Tyr Ile Arg Lys His
Sekundaarsed	Arg Lys	Asp Glu	Asp Glu Cys
Tertsiaarsed	-	Asn Gln	Asn Gln
Stabiliseerivad	His Ile Asn Gln Asp Glu Cys Ala Ser Thr Gly Val Pro Met	Cys Ala Ser Thr Gly Val Pro Met	Ser Gly Val Pro Met

3. Translatsiooni initsiatsiooni mõjutavad järjestuse motiivid

3.1 Bakterid

Bakterites seondub väike ribosomaalne subühik kompleksis mitmete initsiatsioonifaktoritega otse translatsiooni initsiatsiooni regioonile.

Lugedes initsiatsioonikoodoni esimeseks positsiooniks +1, paiknevad elemendid, mis on olulised prokarüootse mRNA translatsiooni initsiatsiooniregiooni äratundmiseks, positsioonide -20 ja +15 vahel (Stormo jt., 1982): translatsiooni initsiatsiooniregiooni sekundaarstruktuur, SD järjestus, enhanser, vahemaa SD ja startkoodoni vahel, startkoodonile vahetult eelnev nukleotiid, startkoodonile järgnev regioon (DR) ja startkoodon ise (joonis 1). Eelistatuimaks startkoodoniks bakterites on AUG seoses tema komplementaarsusega initsiaator-tRNA antikoodoniga. *E. coli*'s kasutatakse järgmisi startkoodoneid: AUG (~82%), GUG (~14%), UUG (~3%), CUG (~0,2%) ja AUU (~0,2%) (Blattner jt., 1997).



Joonis 1. Translatsiooni initsiatsiooniregiooni elemendid bakterites. SD – Shine-Dalgarno järjestus, DR – allavoolu asuv regioon.

3.1.1 Startkoodonist ülesvoolu asuvad motiivid

a) Shine-Dalgarno järjestus on startkoodonile eelnev lühike puriinirikas järjestus (*E. coli*'s AGGAGG), mis paardub komplementaarse järjestusega (anti-SD) ribosoomi väikese subühiku 16S rRNA 3' otsa lähedal (Shine ja Dalgarno, 1974; Steitz ja Jakes, 1975; Sakai jt., 2001). Seondumise peamine tähtsus seisneb ribosoomi paigutamises õige startkoodoni juurde. Pikendatud SD järjestus on palju vähem efektiivne kui lühemad SD järjestused (efektiivseim 6 nt) (de Smit ja van Duin, 1994; Komarova jt., 2002). SD järjestus asub tavaliselt umbes 5-8 nt startkoodonist ülesvoolu, optimaalseimaks vaheks startkoodoni ja SD järjestuse vahel on pakutud 5 nt-i (Chen jt., 1994), mis vastab ka 16S rRNA 3' otsa ja ribosoomi P-saidi vahelise vahemaa pikkusele.

b) Enhanser on A/U rikas järjestus enne SD järjestust. Enhanser seondub mRNA-le seonduva ribosomaalse valguga S1-ga (Boni jt., 1991). S1 on suurim ribosomaalne valk, mis paikneb 30S subühikus. S1 tagab translatsiooni ka sellistelt mRNA-delt, millel puudub või on nõrk, sh ka pikk, SD järjestus (Roberts ja Rabinowitz, 1989; Farwell jt., 1992; Tzareva jt., 1994; Komarova jt., 2002; Komarova jt., 2005).

c) startkoodonile vahetult eelneva nukleotiidi ja tRNA^{Met} 37. positsiooni vahel moodustuv aluspaar soodustab translatsiooni. Kuigi see 5' suunas pikendatud koodon-antikoodon interaktsioon ei ole initsiatsiooni jaoks absoluutselt vajalik, võib see anda lisaspetsiifilisuse mittestandardsete initsiatsioonikoodonitega transkriptidele ja/või säilitada translatsiooni initsiatsiooni spetsiifilisust teatud keskkonna poolt põhjustatud stressitingimustes või juhul, kui mRNA või tRNA hulgad on limiteeritud (Esposito jt., 2003).

3.1.2 Startkoodonist allavoolu asuvad motiivid

Algselt arvati, et startkoodonist allavoolu asuv regioon mõjutab translatsiooni initsiatsiooni mRNA-tRNA komplementaarse aluspaardumise kaudu. Mutatsiooniliste muutuste sisestamisel nii mRNA-sse kui 16S rRNA-sse selline teooria aga kinnitust ei leidnud

(Firpo ja Dahlberg, 1998; O'Connor jt., 1999). Seleksiooniline surve efektiivseks geeniekspressiooniks toob kaasa avatud lugemisraamide järjestuse mittejuhusliku koodonkasutuse (Chen ja Inouye, 1990; Ohno jt., 2001). Erinevate +2 koodonite kasutusel on saavutatud 15-20 kordne translatsiooni efektiivsuse muutus; *E. coli*'s on kõige levinum ja kõige rohkem ekspressioonitaset mõjutav koodon AAA (Stenstrom jt., 2001). Üleüldine G nukleotiidiga algavate koodonite eelistus *E. coli*'s on samuti positiivses korrelatsioonis geenide ekspressioonitasemega (Gutierrez jt., 1996). Samas vähendavad NGG koodonid märkimisväärselt geeniekspressiooni (Gonzalez de Valdivia ja Isaksson, 2004). Siiski pole efekt limiteeritud vaid +2 koodoniga. Kuni 5 koodoni pikkuste erinevate DR-ide uurimisel on selgunud, et nn. tugev DR regioon võib anda translatsiooni tasemel kõrge efekti ka ilma tugeva SD järjestuseta (Stenstrom jt., 2001). A nukleotiidi eelistus esineb *E. coli* geenide alguses 20-30 nukleotiidi ulatuses (Rocha jt., 1999). Kõik 16S rRNA-de üksikahelalised regioonid omavad väga kõrget A-sisaldust (Gutell jt., 1985; Wang ja Hickey, 2002) hoolimata erinevast genoomsest GC-sisaldusest (Gutell jt., 1985). Seetõttu arvatakse, et translatsiooni initsiatsioonil on A-rikas mRNA eelistatud, kuna ta ei sisalda sekundaarstruktuure (Eyre-Walker ja Bulmer, 1993; Rocha jt., 1999; Stenstrom ja Isaksson, 2002).

3.2 Eukarüooidid

Eukarüootide translatsiooni initsiatsioon sõltub m^7G *cap* struktuurist, AUG koodonit ümbritseva primaarjärjestuse kontekstist, AUG asukohast 5' *cap*'i suhtes, liiderjärjestuse pikkusest ja sekundaarstruktuurist AUG koodonist nii üles- kui allavoolu. Erinevalt prokarüootidest on eukarüootides alternatiivsed startkoodonid harvemad (Kozak, 1991). AUG koodoni kaugus 5' otsast on sama oluline kui kontekst; selgroogsete mRNA-dest vähemalt 90% puhul initseerib ribosoom esimeselt AUG koodonilt (Kozak, 1987). Mõned ribosoomid aga võivad mööduda AUG-st, mis asub nõrgas kontekstis (Kozak, 1991). Esimese AUG koodoni äratundmine võib olla näiteks takistatud, kui see asub liiga lähedal

cap struktuurile (esimese 12 nukleotiidi sees alates *cap* struktuurist) (Dabrowski ja Alwine, 1988; Sedman jt., 1990; Gillman jt., 1991). Lisaks võib *cap* struktuurist alates esimese 12 nukleotiidi juures asuv sekundaarstruktuur translatsiooni drastiliselt inhibeerida, kuna takistab 40S ribosoomi seondumist mRNA-le (Kozak, 1989).

3.2.1 Startkoodonit ümbritseva järjestuse kontekst

Ka *S.cerevisiae* puhul väheneb translatsiooni initsiatsiooni efektiivsus, kui startkoodoni kontekst erineb märgatavalt kindlast eelistatud nukleotiidsest kontekstist AUGUC(U/C) (Hamilton jt., 1987; Miyasaka, 1999; Fuglsang, 2004). Nii *S. cerevisiae*'s kui ka *Schizosaccharomyces pombe*'s on levinuimaks teiseks koodoniks UCU (Niimura jt., 2003). Pärsi kõrge ekspressioonitasemega geenides esineb startkoodonist ülesvoolu tugev A-eelistus (Cigan ja Donahue, 1987; Cavener ja Ray, 1991; Yoon ja Donahue, 1992).

Imetajate translatsiooni initsiatsiooniks on optimaalseim kontekst GCCRCCAAUGG, kus on olulisimateks positsioonideks -3 puriin, eelistatult A, ning +4 positsioonis asuv G (Kozak, 1997). Inimesel on levinuimaks startkoodonile järgnevaks koodoniks GCG (Niimura jt., 2003).

3.3 Arhead

Arheate SD järjestuse konsensusjärjestus sõltub oluliselt vastava 16S rRNA 3' otsa järjestusest, mille pikkus erinevates liikides varieerub (Torarinsson jt., 2005). Arheate ennustatud kõrge ekspressioonitasemega geenide ja SD motiivi tugevuse vahel on näidatud positiivne korrelatsioon (Karlin jt., 2005). Ka startkoodoni kasutus varieerub erinevates liikides märkimisväärselt, kuigi kõige levinuimaks startkoodoniks on ka arheates AUG, mis moodustab erinevates organismides 70-90% kõigist ennustatud startkoodonitest. Samas näiteks *Aeropyrum pernix*'i, *Hyperthermus butylicus*'e ja *Methanopyrus kandleri* puhul

kasutatakse AUG, GUG ja UUG startkoodoneid võrdselt; *Pyrobaculum aerophilum* aga kasutab eelistatult GUG koodonit (32%) (Torarinsson jt., 2005).

4. Koodonkasutus ja geenide ekspressioonitase

Organismid kasutavad sünonüümseid koodoneid väga erineva sagedusega. Koodoneelistus on märkimisväärselt tugevam kõrge ekspressioonitasemega geenides ning on näidatud selge korrelatsioon sünonüümsete koodonite kasutuse ja valgu hulga vahel (Ikemura, 1985). Selline eelistuse erinevus on translatsiooni efektiivsusele suunatud selektsiooni tagajärg. Kõrge ekspressioonitasemega geenide spetsiifilise koodoneelistuse tekke peamiseks põhjuseks on mitmete autorite poolt pakutud tRNA-de kättesaadavuse varieerumine (näit (Andersson ja Kurland, 1990)).

4.1 Koodoneelistuse mõõtmise meetodid

Kõik koodonkasutusel põhinevad geeni ekspressioonitaseme ennustamise meetodid lähtuvad referentsandmestikust, mis koosneb eksperimentaalselt kindlaks tehtud kõrge ekspressioonitasemega geenidest. Seega mõõdavad sellised meetodid kõrge ekspressioonitasemega geenide põhjal leitud nn. optimaalsete koodonite kasutuse ulatust geenis (näiteks F_{op} (Ikemura, 1981), CBI (Bennetzen ja Hall, 1982), ENC (Wright, 1990)). Viimasel ajal on sagedamini kasutatud leidnud koodoni adaptatsiooni indeks (CAI) ja koodonkasutuse erinevuste indeks.

4.1.1 CAI

CAI on meetod, mis iseloomustab, kui sarnane on antud geeni koodonkasutus kõrge ekspressioonitasemega geenide koodonkasutusele (Sharp ja Li, 1987). Algselt loodi see

S.cerevisiae põhjal, kuid on kasutatav ka teistes organismides. CAI omistab igale koodonile (v.a stoppkoodonid) parameetri, milleks on relatiivne adaptiivsus (w). Koodoni relatiivne adaptiivsus on defineeritud kui tema sageduse suhe kõige rohkem kasutatavasse sünonüümsesse koodonisse. Parameetri arvutamise aluseks on kõrge ekspressioonitasemega geenide grupp, kuhu kuuluvad ribosoomivalke, välismembraanivalke, elongatsioonifaktoreid, kuumaehmatuse valke ja RNA polümeraasi subühikuid kodeerivad geenid. Relatiivne adaptiivsus väljendub järgnevalt:

$$w_{aa,i}(G) = \frac{f_{aa,i}(G)}{f_{aa,max}(G)}, \quad (1)$$

kus $f_{aa,i}$ on koodoni i (mis kodeerib aminohapet aa) sagedus ja $f_{aa,max}$ on kõige rohkem kõrge ekspressioonitasemega geenides G aminohappe aa kodeerimiseks kasutatava koodoni sagedus. Relatiivne adaptiivsus varieerub 0-st 1-ni – vastavalt 0, kui koodonit ei kasutata rühmas G üldse, ja 1, kui koodon esineb antud aminohappe puhul rühmas G kõige rohkem. Geeni g CAI on kõigi geenijärjestuses olevate koodonite relatiivsete adaptiivsuste geomeetriline keskmine:

$$CAI_g = \prod_{i=1}^N w_i^{1/N}, \quad (2)$$

kus w_i on i -nda koodoni relatiivne adaptiivsus geenis, milles on N koodonit. Valemi (2) saab muuta järgnevaks valemiks:

$$CAI_g = \prod_{k=1}^{61} w_k^{X_{k,g}}, \quad (3)$$

kus w_k tähistab nüüd geneetilise koodi 61 koodonist k -nda koodoni relatiivset adaptiivsust. $X_{k,g}$ on koodoni k fraktsioon geenis g olevate koodonite koguarv:

$$X_{k,g} = \frac{C_{k,g}}{\sum_{i=1}^{61} C_{i,g}}, \quad (4)$$

kus $C_{k,g}$ on koodoni k esinemise arv geenis g . Nagu relatiivne adaptiivsusi, varieerub CAI 0-st 1-ni. Kõrgemad CAI väärtused näitavad, et vastavad geenid on tõenäoliselt kõrgema ekspressioonitasemega (Sharp ja Li, 1987).

Pärmi puhul algselt kasutatud vähene andmehulk (24 geeni) oli ettenägelikult valitud ning andis adekvaatse tulemuse; hoolimata suurest hulgast praegu kättesaadavatest ekspressioonandmetest ei ole õnnestunud hiljem korrelatsiooni pärmi CAI ja tegeliku ekspressioonitaseme vahel tõsta rohkem kui 0,01 punkti võrra (Jansen jt., 2003). Lisaks on näidatud tugev positiivne korrelatsioon CAI ja pärmi SAGE analüüsist saadud ekspressioonandmete vahel (Futcher jt., 1999). SAGE meetod põhineb igale transkriptile vastavate unikaalsete järjestuse *tag*'ide esinemise kvantitatiivsel analüüsil (Velculescu jt., 1995; Velculescu jt., 1997). Erinevatest olemasolevatest sünonüümse koodonkasutuse indeksitest annab samuti just CAI kõige tugevama positiivse korrelatsiooni *S. cerevisiae* SAGE andmetega (Coghlan ja Wolfe, 2000). Ka *E. coli* puhul on näidatud tugev korrelatsioon CAI ja eksperimentaalsete geeniekspressiooni andmete vahel (dos Reis jt., 2003; Jia ja Li, 2005).

4.1.2 Koodonkasutuse erinevuste indeks

Koodonkasutuse erinevuste indeks avaldati kui alternatiivne võimalus CAI-le ning see hindab ühe grupi geenide (või ühe geeni) koodonkasutuse erinevust teise grupi geenide suhtes (Karlin ja Mrazek, 2000). Geeni *g* ekspressioonimõõtu $E(g)$ defineerub järgmiselt:

$$E(g) = \frac{B(g|C)}{\frac{1}{2}B(g|RP) + \frac{1}{4}B(g|Ch) + \frac{1}{4}B(g|Tf)}, \quad (5)$$

kus geenikomplekt *C* sisaldab kõiki genee genoomis, *RP* ribosomaalseid valke, *Ch* *chaperon*'e ja *Tf* transkriptsioonifaktoreid. Kõrgelt ekspresseeruvad geenid omavad vastavalt kõrgemat $E(g)$ väärtust kui madalalt ekspresseeruvad geenid (Karlin ja Mrazek, 2000). Selle indeksi puhul on miinuseks asjaolu, et ta ülehindab lühikeste (kuni 500-koodoniliste) geenide koodoneelistust (Supek ja Vlahovicek, 2005).

4.2 Koodonkasutuse ja peptidüül-tRNA *drop-off* i seos

Peptidüül-tRNA-de enneaegne ribosoomist vabanemine ehk *drop-off* on normaalne sündmus, mis võib toimuda elongatsiooni käigus (Dincbas jt., 1999; Menez jt., 2000).

Juhul kui *drop-off* reaktsioon on väga efektiivne, küllastub peptidüül-tRNA-de retsükleerimise eest vastutav ensüüm peptidüül-tRNA hüdrolaas. Seetõttu akumulereuvad tRNA'd peptidüül-tRNA vormis rakku ning selle tulemusena tekkiv deaminoatsüleeritud tRNA-de puudujääk takistab efektiivset translatsiooni (Hernandez-Sanchez jt., 1998; Dincbas jt., 1999; Tenson jt., 1999; Heurgue-Hamard jt., 2000; Menez jt., 2000). *Drop-off* reaktsiooni kiirus sõltub sünteesitud peptiidahela pikkusest ja koodonist. Mida lühem on sünteesitud peptiidahel, seda efektiivsem on *drop-off* (Heurgue-Hamard jt., 2000). Kõige tõenäolisemalt toimub selliste tRNA-de *drop-off*, mis loevad koodoneid, mille esimeseks ja/või teiseks nukleotiidiks on A (Cruz-Vera jt., 2003). Järelikult võiksid A-rikkad koodonid ORF-ide alguses põhjustada kõrge sagedusega *drop-off* i toimumist.

PRAKTILINE OSA

TÖÖ EESMÄRK

Kuigi translatsiooni initsiatsiooniregiooni järjestuse konteksti on eelnevalt palju uuritud, põhinevad tehtud tööd tavaliselt kõigil uuritava organismi geenidel. Käesolevas töös analüüsiti väiksemat gruppi kõrge ekspressioonitasemega geene erinevates organismides eesmärgiga leida neile iseloomulikke ja organismi teistest geenidest erinevaid signaale avatud lugemisraamide alguses.

ANDMED JA MEETODID

1. Andmed

Analüüsitavate organismide hulka valisime levinud mudelorganismid (*E. coli* ja *S. cerevisiae*), mille kohta on teada ka geeniekspressiooni andmed, ning lisaks võimalikult erinevate omadustega organismid: väikese, keskmise ja suure genoomiga, erineva genoomse GC-sisaldusega, vabalt elavad organismid ja obligatoorsed parasiidid, ekstremofiilid. Igast genoomist moodustasime kaks erinevat andmestikku: kõrge ekspressioonitasemega geene sisaldav andmestik ja vastava organismi kõiki geene sisaldav andmestik.

Kõikide analüüsitud üherakuliste organismide valke kodeerivad järjestused pärinevad GenBank andmebaasist: *Escherichia coli* K12 [NC_000913], *Bacillus subtilis* [NC_000964], *Haemophilus influenzae* [NC_000907], *Helicobacter pylori* 26695 [NC_000915], *Mycobacterium tuberculosis* H37Rv [NC_000962], *Treponema pallidum* [NC_000919], *Rickettsia prowazekii* [NC_000963], *Mycoplasma genitalium* [NC_000908], *Borrelia burgdorferi* [NC_001318], *Mycobacterium avium* paratuberculosis [NC_002944], *Synechococcus elongatus* PCC 7942 [NC_007595], *Rickettsia conorii* [NC_003103], *Methanococcus jannaschii* [NC_000909], *Archaeoglobus fulgidus* [NC_000917], *Pyrococcus horikoshii* [NC_000961], *Haloarcula marismortui* [NC_006396 –

NC_006397], *Saccharomyces cerevisiae* [NC_001133 – NC_001148], *Schizosaccharomyces pombe* [NC_003421, NC_003423 – NC_003424], *Plasmodium falciparum* [NC_000521, NC_000910, NC_004314 – NC_004318, NC_004325 – NC_004331], *Cryptococcus neoformans* [NC_006670, NC_006679 – NC_006687, NC_006691 – NC_006694].

Homo sapiens'i valke kodeerivate geenide järjestused pärinevad UCSC *Human Genome Browser*'ist (Kent jt., 2002; Karolchik jt., 2003) ja koosnevad CCDS järjestustest (*consensus coding sequence*) (14814 tk). CCDS järjestused on järjestused, mille kodeeriva järjestuse genoomsed koordinaadid on NCBI RefSeq, ENSEMBL ja Vega andmebaaside geeniannotatsioonides identsed, omavad korrektset start- ja stoppkoodonit, on genoomsest järjestusest transleeritavad ilma raaminihketa ning omavad konsensus splaissingusaite ja homoloogiat teistes liikides (Karolchik jt., 2003).

1.1 Kõrge ekspressioonitasemega geenid üherakulistest organismides

Kõrge ekspressioonitasemega geenide (KEG) andmehulkade koostamisel võtsime aluseks CAI indeksi, kuna CAI ja geeniekspressiooni taseme vahel esineb tugev korrelatsioon (Futcher jt., 1999; Coghlan ja Wolfe, 2000; dos Reis jt., 2003; Jansen jt., 2003).

E. coli ja *S. cerevisiae* KEG andmestiku jaoks arvutasime kõigile genoomi geenidele programmi CodonW (John Peden, versioon 1.3, <http://www.molbiol.ox.ac.uk/cu>) abil CAI ning valisime 80 kõrgeima CAI väärtusega geeni (lisa I,II).

Ülejäänud organismide KEG andmestike koostamine põhines eeldusel, et funktsiooni konserveerumine viitab ka suhtelisele geeni ekspressiooni taseme konserveerumisele. Seetõttu koosnesid teiste analüüsitud bakterite KEG andmestikud 80-le *E. coli* KEG-ile vastavatest ortoloogidest¹ (lisa I), *S. pombe* ja *P. falciparum*'i KEG andmestik sisaldas 80-le *S.cerevisiae* KEG-le vastavaid ortolooge (lisa II) ning arheate KEG andmehulk sisaldas nii *E. coli* kui ka *S. cerevisiae* 80 KEG-i ortolooge (lisa I, II). Ortoloogid leidsime kahe

¹ Ortoloogid on geenid kahest eri liigist, mis pärinevad ühest geenist nende liikide viimases ühises eellases.

genoomi võrdlusel BLAST-iga (Altschul jt., 1997) kõik valgud kõigi vastu kasutades programmi InParanoid (Remm jt., 2001) ja valisime sellised paarid, mis andsid mõlemas suunas parima skoori.

1.2 Kõrge ekspresioonitasemega geenid inimeses

Inimese kõrgelt ekspresseerunud geenide defineerimise aluseks olid SAGE andmed 24 erineva koe 84 raamatukogust. SAGE meetod põhineb igale transkriptile vastavate unikaalsete järjestuse *tag*'ide esinemise kvantitatiivsel analüüsil (Velculescu jt., 1995; Velculescu jt., 1997). SAGE *tag*'id on vastavusse seatud UniGene geeniklastritega, millest igaüks sisaldab järjestusi, mis esindavad unikaalset geeni (Wheeler jt., 2003). Käesolevas töös teostatavate analüüside puhul on äärmiselt oluline kasutada ainult korrektse alguskoordinaadiga valku kodeerivaid järjestusi. Seetõttu võrdlesime kõiki vastavaid UniGene'i klastrite järjestusi CCDS järjestustega BLAST-i abil ning valisime potentsiaalsesse KEG andmehulka ainult parima skoori andnud CCDS järjestused. Pärast korduvate järjestuste välja sorteerimist sisaldas lõplik inimese KEG andmestik 975 geeni, mille hulgas olid ka koetsiifilised geenid.

2. Meetodid

2.1 A nukleotiidi sageduse ja CAI seose analüüs

KEG-ide alguse A nukleotiidi sageduse ja geeniekspressiooni vahelise seose kontrollimiseks jagasime kõik *E. coli* geenid CAI alusel seitsmesse gruppi. Igale grupile arvutasime keskmise A sageduse koodonites 2 ja 3-5.

2.2 Nukleotiidide eelistuse analüüs

E. coli geenide alguse adensiini kasutuse analüüsiks arvutasime igale koodonile esimese kümne koodoni hulgas nii KEG-ides kui kõigis geenides keskmise A sageduse.

Lisaks võrdlesime nukleotiidide sagedusi KEG-ide esimese kümne koodoni kõigis kolmes nukleotiidi positsioonis eraldi kõigi geenide vastavates positsioonides asuvate nukleotiidide sagedustega. H_0 : KEG-ide ja kõigi geenide nukleotiidide sageduste vahel puudub erinevus.

G_4 ja C_5 nukleotiidide suurenenud sageduste omavahelise sõltuvuse või sõltumatuse kontrolliks võrdlesime nende nukleotiidide sagedusi KEG ja kõigi geenide andmestikes. G_4 sageduste võrdlusel eemaldasime mõlemast andmehulgast geenid, mille teiseks koodoniks oli NCN koodon. C_5 sageduste võrdlusel eemaldasime teise koodonina GNN koodonit kasutavad geenid. H_0 : G_4 ja C_5 nukleotiidide väljaspoolalaniini koodoneid esinemise sagedused ei erine KEG ja kõigi geenide andmestikes.

2.3 Koodonite eelistuse analüüs

Koodonite üle- ja alaesindatuse leidmiseks võrdlesime KEG-ide alguse koodonkasutust kõigi geenide alguse koodonkasutusega eraldi igas koodonipositsioonis. H_0 : koodonite sagedused ei erine KEG-ides ja kõigis geenides.

Lisaks analüüsisime spetsiifilisemaltalaniini koodonite (GCN) kasutust KEG-ides. Selleks võrdlesime erinevate GCN koodonite kasutust KEG-ide teises koodoni positsioonis ja kogu KEG-ide ulatuses. H_0 :alaniini koodonite sagedused ei erine KEG-ide teises koodonis ja kõigis koodonites kogu KEG-ide ulatuses.

2.4 Aminohapete eelistuse analüüs

Sarnaselt koodonite üle- ja alaesindatuse leidmise analüüsile võrdlesime ka aminohapete puhul kõrgelt ekspresseeruvate valkude alguse aminohappe kasutust kõigi valkude alguse

aminohappe kasutusega. H_0 : aminohapete sagedused ei erine kõrgelt ekspresseeruvates valkudes ja kõigis valkudes.

Eraldi analüüsisime nii eukarüootides kui bakterites valkude stabiilsust ning fMet lõikamist soodustavate aminohapete (Ala, Gly, Pro, Ser, Thr ja Val) eelistust kõrgelt ekspresseeruvate valkude teises aminohappe positsioonis. H_0 : selle grupi aminohapete sagedused ei erine kõrgelt ekspresseeruvate valkude ja kõigi valkude teises positsioonis.

3. Statistiline olulisus

Nukleotiidide, koodonite ja aminohapete sageduste võrdlemiseks KEG andmestikus ja kõigi geenide andmestikus kasutasime Fisher'i täpset testi (FET). Sarnaselt χ^2 -testile näitab ka FET, kas uuritavate sündmuste sagedused kahes erinevas andmehulgas on piisavalt erinevad, et lükata ümber nullhüpotees. Kuigi χ^2 -testi on lihtsam arvutada, töötab FET paremini väikeste andmehulkade puhul (kui arvud sagedustabelis on väiksemad kui 6). Lisaks on FET-iga saadud p-väärtus täpne, samas kui χ^2 -test annab tulemuseks p-väärtuste vahemiku.

Kõikides analüüsides lugesime statistiliselt oluliseks tulemused, mille p-väärtus oli võrdne või väiksem kui 0,01. Korrektsiooni mitmese testimise mõju vähendamiseks ei ole üheski testis kasutatud.

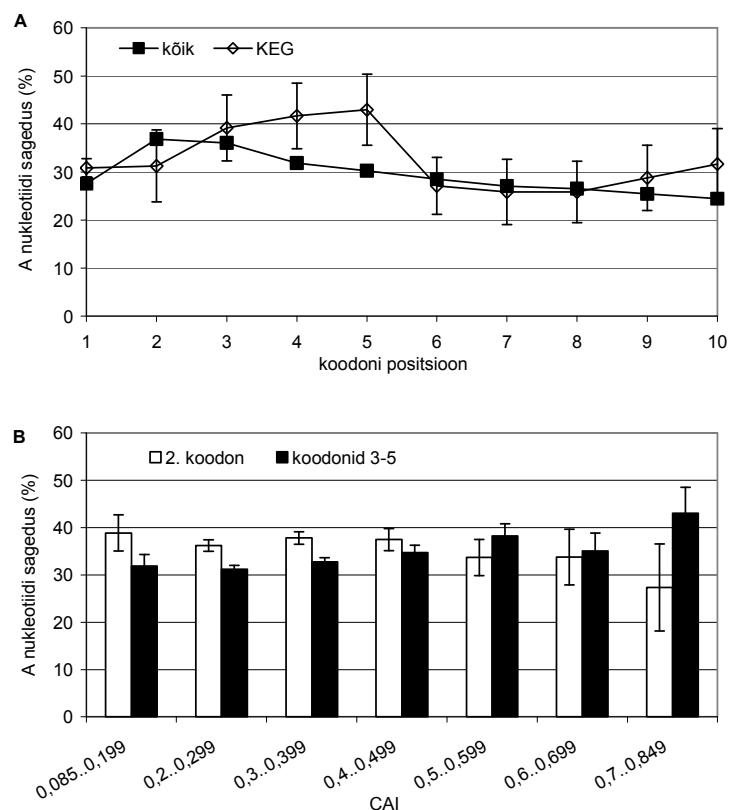
TULEMUSED JA ARUTELU

1. Adenosiini sagedus *E. coli* avatud lugemisraamide alguses

Esialgseks järjestuste eelistuste analüüsimiseks *E. coli* ORF-ide alguses lugesime kokku nukleotiidide sagedused iga koodonipositsiooni kohta. Kõigi *E. coli* geenide analüüs näitas, et ORF-ide alguses on tõusnud adenosiini sagedus (joonis 2A). Ka eelnevalt on näidatud, et *E. coli*'s eelistatakse startkoodoni järel A-rikkaid järjestusi (Rocha jt., 1999). Arvatakse, et selline eelistus tuleb vajadusest vähendada võimalike mRNA sekundaarstruktuuride stabiilsust initsiatsiooniregioonis (Rocha jt., 1999; Stenstrom jt., 2001).

Samuti on eelnevalt teada, et kõrge ekspressioonitasemega geenid omavad väga erinevat koodonkasutust ja seega ka nukleotiidsel eelistuse mustrit. Seetõttu analüüsisime A sagedust ka *E. coli* KEG andmestikus, mis koosnes 80-st kõrgeima CAI-ga geenist. Selgus, et A sagedus koodonites 3-5 oli umbes 1,3 korda kõrgem KEG-ides võrreldes kõigi geenide andmehulgaga ($p = 2,2E-06$). Samas ei olnud A nukleotiidi sagedus KEG-ide teises koodonis tõusnud, vaid hoopis *ca* 1,3 korda langenud võrreldes kõigi geenide vastavate positsioonidega, kuigi statistiline olulisus oli madal ($p = 0,079$) (joonis 2A).

Selleks, et veenduda, et A nukleotiidide sageduse erinevus koodonites 2 ja 3-5 on seotud geenide ekspressioonitasemega, teostasime järgneva analüüsi: jagasime kõik *E. coli* geenid CAI väärtuste alusel seitsmesse gruppi ning võrdlesime A kasutust nendes gruppides. Selgus, et A nukleotiidi eelistus teises koodonis vähenes ainult kõige kõrgema ekspressioonitasemega geenide grupis. CAI väärtuste ja A nukleotiidi sageduste vahel koodonites 3-5 esines aga positiivne seos (joonis 2B).

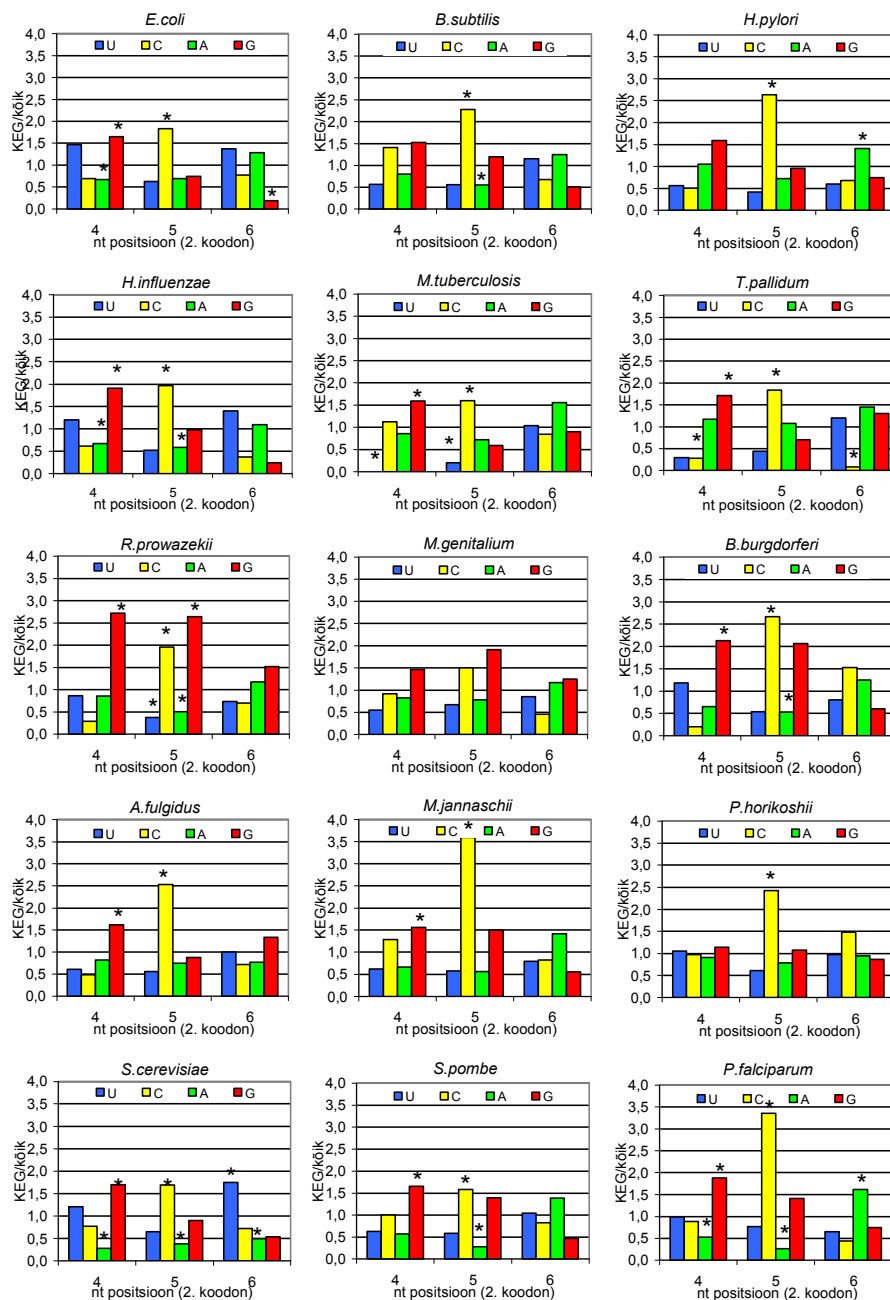


Joonis 2. A nukleotiidi sagedus *E.coli* ORF-ide alguses. Näidatud on keskmine sagedus koodoni kohta. **A.** Kõrgelt ekspresseerunud geenides on statistiliselt oluline A sageduse tõus koodonites 3-5 (nukleotiidid 7-15), kuid langus teises koodonis võrreldes kõigi geenidega. **B.** A nukleotiidi eelistus teises koodonis (nukleotiidid 4-6) väheneb ekspressioonitaseme tõustes. Ekspressioonitaseme ja A nukleotiidi sageduse vahel koodonites 3-5 esineb aga positiivne korrelatsioon. Veapiirid näitavad 1,96-kordset standardviga.

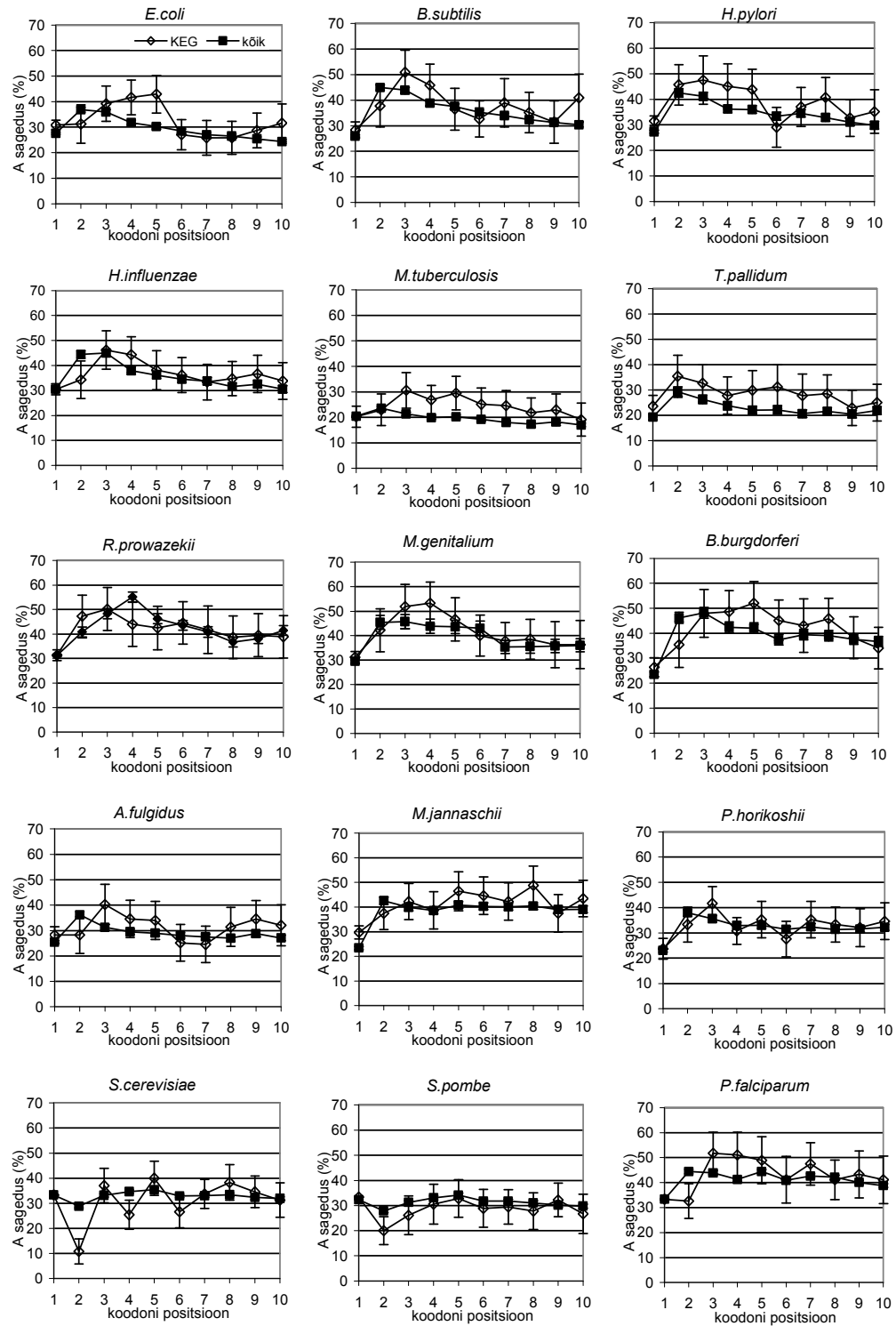
2. Nukleotiidide kasutus erinevate üherakuliste organismide ORF-ide alguses

A nukleotiidi sageduse langus *E. coli* KEG-ide teises koodonis viitab sellele, et kõrge ekspressioonitasemega geenidel esinevad tõepoolest erinevad järjestuse eelistused võrreldes kõigi geenidega. Tekkisid küsimused, kas selline muster on ainult *E. coli* spetsiifiline või

laiemalt levinud ning kas nähtud A nukleotiidi langus võib olla hoopis teiste nukleotiidide sageduste muutuste tulemus. Nendele küsimustele vastamiseks võtsime vaatluse alla lisaks *E. coli*'le veel 8 bakterit, 3 arheat ja 3 üherakulist eukarüooti ning analüüsisime nendes genoomides kõigi erinevate nukleotiidide kasutust initsiatsioonikoodonist allavoolu. Iga genoomi puhul võrdlesime omavahel vastavaid nukleotiidipositsioone KEG-ide andmestikus ja kõigi geenide andmestikus. 14 genoomis 15-st analüüsitust oli tõusnud C sagedus KEG-ide viiendas nukleotiidipositsioonis (vastab teise koodoni teisele nukleotiidile, $p < 0,01$) (joonis 3). Ainsana ei olnud C_5 sageduse tõus statistiliselt oluline *M. genitalium*'is. Lisaks oli KEG-ides tõusnud ka G_4 nukleotiidi sagedus võrreldes kõigi geenidega, kuigi mitte nii laialdaselt kui C_5 sagedus. Tõus oli statistiliselt oluline 11 genoomis 15-st analüüsitust ($p < 0,01$). Enamusel juhtudel kaasnes G_4 ja C_5 sageduste suurenemisega tõepoolest A nukleotiidi sageduse langus, kuid esines ka U nukleotiidi sageduse langust (joonis 3, lisa III). *E. coli*'s nähtud A sageduse tõus KEG-ide koodonites 3-5 (nukleotiidid 7-15) esines ka *M. tuberculosis*'e KEG-ide puhul ($p = 2,7E-06$) (joonis 4). Kuigi nendes kahes genoomis on tendents tugev, ei ole see laiemaalt konserveerunud.



Joonis 3. Nukleotiidide kasutus KEG-ide teises koodonis. Nukleotiidide sagedused KEG-ide teise koodoni kõigis kolmes positsioonis on jagatud vastavate sagedustega kõigis geenides. Tärnid tähistavad 0,01-st väiksemaid p-väärtuseid (H_0 : nukleotiidide sageduste vahel KEG-ides ja kõigis geenides puudub erinevus).



Joonis 4. A nukleotiidi sagedused KEG-ide ja kõigi geenide esimeses 10 koodonis.

3. Koodonite eelistused üherakulistes organismides

Ilmnenud nukleotiidide kasutuse muster viitab võimalikule GCN koodoni eelistusele KEG-ide teise koodonina. Seetõttu võrdlesime ka koodonite kasutust üherakuliste organismide KEG-ide alguses ja kõigi geenide alguses (tabel 2).

Tabel 2. Koodoneelistused KEG-ide alguses võrreldes kõigi geenidega (H_0 : koodonite sagedused KEG-ides ja kõigis geenides ei erine).

organism	koodoni positsioon											
	2			3			4			5		
	koodon	p-väärtus	%KEG/ %kõik	koodon	p-väärtus	%KEG/ %kõik	koodon	p-väärtus	%KEG/ %kõik	koodon	p-väärtus	%KEG/ %kõik
<i>E.coli</i>	GCU	7,9E-06	15/3	ACU	1,1E-04	10/2	AUU	4,0E-05	16/4	AAA	0,001	15/5
	GCA	6,0E-04	11/3	AAG	0,004	11/4	ACU	0,007	6/1			
	UCC	7,1E-04	8/1									
<i>B.subtilis</i>	GCA	8,5E-05	15/3	GUA	0,004	9/2	GGA	0,002	9/2		-	
<i>H.pylori</i>	GCA	0,001	15/4		-			-		AAG	0,009	9/2
	ACA	0,005	6/1									
<i>H.influenzae</i>	GCA	1,0E-04	17/4	GUA	0,005	6/1	GAA	0,009	11/4		-	
	GCU	0,002	11/3									
	UCU	0,006	10/3									
<i>M.tuberculosis</i>	GCA	3,1E-04	11/2	AAG	2,9E-04	11/2	AAG	2,1E-04	11/2	AAG	8,5E-05	11/2
	GCG	0,008	11/4							ACU	0,008	5/1
<i>T.pallidum</i>	GCA	6,9E-04	15/3	AAG	0,010	13/4		-			-	
<i>R.prowazekii</i>		-			-			-			-	
<i>M.genitalium</i>		-			-			-			-	
<i>B.burgdorferi</i>	GCA	1,0E-04	15/2		-			-			-	
	UCA	9,9E-04	10/1									
	GGU	0,004	8/1									
<i>A.fulgidus</i>		-			-		AAG	0,004	15/5	AAG	0,002	15/4
										UAU	0,006	8/1
										AGA	0,006	11/3
<i>M.jannaschii</i>	GCA	1,1E-05	14/2	GGA	0,003	9/2		-			-	
	GCU	0,002	9/2									
	CCA	0,008	7/1									
<i>P.horikoshii</i>		-			-		AUG	0,004	27/12		-	
<i>S.cerevisiae</i>	GCU	7,3E-08	19/3	AGA	2,5E-07	18/3	GUU	8,6E-08	15/2	AAG	8,4E-04	13/4
	UCU	7,2E-07	24/6	CCA	0,006	6/1	GGU	0,002	8/2	ACU	0,001	11/3
	GGU	1,0E-05	13/2				CCA	0,005	8/2	GUU	0,002	9/2
	GCC	0,006	6/1									
<i>S.pombe</i>	GCA	2,0E-07	22/4	CGU	4,5E-06	13/1		-		AAG	8,7E-04	13/3
	GGA	0,002	8/1	AUU	0,001	13/3				AUC	0,009	7/1
<i>P.falciparum</i>	GGA	0,001	11/2		-		CAA	0,004	11/2		-	
	GCA	0,001	11/2									
	UCA	0,002	11/2									
	GCU	0,002	11/2									
	CCA	0,006	6/1									
<i>H.sapiens</i>		-			-		AUC	0,003	3/2	AUC	2,5E-04	3/2
							CUC	0,010	4/2	UGG	0,009	3/2
										UCU	0,001	0/1

Selgus, et 11 organismis 15-st analüüsitust esines ühe GCN koodoni sageduse oluline tõus ($p < 0,01$) KEG-ides teises koodonipositsioonis. Teisele koodonile järgnevates positsioonides ei esinenud sarnast konserveerunud kindlate koodonite eelistust erinevates organismides. Siiski on näha, et tihti eelistatakse positsioonides 3-5 A-rikkaid koodoneid (lüsiini koodonid AAA ja AAG ning arginiini koodon AGA). Võimalikuks põhjuseks võib olla A-rikka ja seega võimalikult vähe sekundaarstruktuure moodustava järjestuse eelistus startkoodonist allavoolu. Siiski jääb selgusetuks, miks on valitud just sellised koodonid kõigist võimalikest A-rikastest koodonitest. Näiteks ei ole ülesindatud asparagiini koodonid (AAU ja AAC). Samas ei olnud ka ükski koodon KEG-ide alguses märkimisväärselt vältitud.

GCN koodonite sageduse tõus teise koodonina võib olla eespool kirjeldatud G_4 ja C_5 nukleotiidide sageduste tõusu tulemus, mitte konkreetsete koodonite eelistus. Sellisel juhul oleksid G_4 ja C_5 ülesindatused üksteisest sõltumatud. Kui tegu oleks aga siiski survega GCN koodonite eelistusele, siis tekitaks see olukorra, kus G_4 nukleotiidi ülesindatus oleks korrelatsioonis C_5 sageduse suurenemisega. Selgitamaks välja, kas tegu on survega nukleotiidsel või koodonite tasemel, teostasime järgnevad analüüsid: kõigepealt eemaldasime andmestikest geenid, mille teiseks koodoniks oli GNN koodon ning testisime, kas C_5 eelistatus KEG-ides võrreldes kõigi geenidega jääb kehtima. Seejärel eemaldasime andmestikest geenid, mille teiseks koodoniks oli NCN koodon ning testisime analoogselt G_4 eelistatuse säilimist. Kerge G_4 ja C_5 ülesindatus küll säilis (tabel 3), kuid see oli palju nõrgem kui GCN koodonite ülesindatus KEG-ide teise koodonina (tabel 2). Koodonanalüüsist ilmneb ka teiste G_4 ja C_5 koodonite eelistus lisaks GCN koodonitele (tabel 2), kuid see eelistus on üldiselt nõrgem ning ei ole erinevate liikide vahel nii ulatuslikult konserveerunud kui GCN koodonite eelistus.

Tabel 3. G₄ ja C₅ nukleotiidide suhe KEG-ide ja kõigi geenide andmestike vahel. G₄ testi jaoks eemaldati andmestikest geenid, mille teiseks koodoniks oli NCN koodon, ja C₅ testi jaoks eemaldati andmestikest geenid, mille teiseks koodoniks oli GNN koodon (H₀: nukleotiidide sagedused ei erine KEG-ides ja kõigis geenides).

organism	G ₄		C ₅	
	p-väärtus	KEG/kõik suhe	p-väärtus	KEG/kõik suhe
<i>E.coli</i>	1,000	1,0	0,048	1,5
<i>B.subtilis</i>	0,490	0,7	0,111	1,7
<i>H.pylori</i>	0,644	1,2	0,004	2,6
<i>H.influenzae</i>	0,336	1,4	0,010	1,8
<i>M.tuberculosis</i>	0,010	0,0	0,495	1,2
<i>T.pallidum</i>	0,524	1,2	1,000	0,9
<i>R.prowazekii</i>	0,012	2,8	0,010	1,9
<i>M.genitalium</i>	0,298	1,3	0,763	1,2
<i>B.burgdorferi</i>	0,011	2,1	0,011	2,6
<i>A.fulgidus</i>	0,177	1,4	0,073	2,0
<i>M.jannaschii</i>	0,282	1,3	0,033	2,5
<i>P.horikoshii</i>	1,000	1,0	0,030	2,2
<i>S.cerevisiae</i>	0,006	1,8	2,8E-04	1,7
<i>S.pombe</i>	0,011	1,7	0,010	1,6
<i>P.falciparum</i>	0,004	2,1	1,3E-05	3,9

Järgnevalt kontrollisime, kas võib olla tegu mõne spetsiifilise GCN koodoni eelistusega KEG-ide teises koodoni positsioonis või toimib surve hoopis aminohapete tasemel. Selleks võrdlesime erinevate GCN koodonite kasutust KEG-ide teises positsioonis ning kogu KEG-ide ulatuses. Kuna erinevate GCN koodonite kasutus ei erinenud märkimisväärselt KEG-ide teises koodonipositsioonis ja ülejäänud KEG-ide positsioonides (tabel 4), viitab see asjaolule, et eelistuse põhjuseks võib olla GCN koodonite poolt kodeeritav aminohapealaniin, mitte mõni spetsiifiline GCN koodon. Üksikute genoomide KEG-ides nähtav GCA koodonite suurem eelistus teise koodonina võib olla seotud suurenenud A nukleotiidi sagedusega koodonites 3-5, mis esines samuti ainult paaris bakteris (joonis 4).

Tabel 4. GCN koodonite sagedused erinevate organismide KEG-ide teises positsioonis ja kogu KEG-ide ulatuses (H_0 : GCN koodonite sagedused ei erine KEG-ide teises positsioonis ja kõigis KEG-ide positsioonides).

organism	koodon	koodoni protsent		p-väärtus	organism	koodon	koodoni protsent		p-väärtus
		KEG-ide 2.pos	kogu KEG				KEG-ide 2.pos	kogu KEG	
<i>E.coli</i>	GCU	57,1	41,8	0,184	<i>A.fulgidus</i>	GCU	21,4	24,4	1,000
	GCC	0,0	8,8	0,250		GCC	21,4	22,7	1,000
	GCA	42,9	25,7	0,082		GCA	28,6	28,0	1,000
	GCG	0,0	23,7	0,007		GCG	28,6	24,8	0,758
<i>B.subtilis</i>	GCU	33,3	40,2	0,792	<i>M.jannaschii</i>	GCU	38,5	48,3	0,582
	GCC	0,0	8,1	0,625		GCC	0,0	4,3	1,000
	GCA	53,3	31,3	0,091		GCA	61,5	44,7	0,268
	GCG	13,3	20,4	0,749		GCG	0,0	2,8	1,000
<i>H.pylori</i>	GCU	28,6	41,3	0,419	<i>P.horikoshii</i>	GCU	0,0	38,0	0,049
	GCC	0,0	16,6	0,144		GCC	42,9	22,6	0,199
	GCA	57,1	11,8	7,3E-05		GCA	28,6	31,0	1,000
	GCG	14,3	30,2	0,251		GCG	28,6	8,4	0,113
<i>H.influenzae</i>	GCU	40,0	26,1	0,199	<i>S.cerevisiae</i>	GCU	68,2	75,7	0,453
	GCC	0,0	4,9	0,621		GCC	22,7	23,4	1,000
	GCA	60,0	50,4	0,502		GCA	9,1	0,6	0,010
	GCG	0,0	18,7	0,037		GCG	0,0	0,3	1,000
<i>M.tuberculosis</i>	GCU	19,2	8,0	0,054	<i>S.pombe</i>	GCU	17,6	62,5	2,2E-04
	GCC	26,9	49,4	0,029		GCC	5,9	31,5	0,031
	GCA	26,9	7,6	0,003		GCA	76,5	5,3	1,4E-13
	GCG	26,9	35,0	0,535		GCG	0,0	0,7	1,000
<i>T.pallidum</i>	GCU	21,4	20,3	1,000	<i>P.falciparum</i>	GCU	45,5	53,6	0,763
	GCC	0,0	12,2	0,397		GCC	0,0	14,8	0,382
	GCA	50,0	24,7	0,055		GCA	45,5	31,1	0,334
	GCG	28,6	42,8	0,416		GCG	9,1	0,4	0,061
<i>R.prowazekii</i>	GCU	44,4	45,4	1,000	<i>H.sapiens</i>	GCU	20,8	26,1	0,071
	GCC	11,1	4,8	0,365		GCC	32,9	43,2	0,002
	GCA	44,4	42,2	1,000		GCA	19,9	20,2	1,000
	GCG	0,0	7,6	1,000		GCG	26,4	10,5	1,6E-11
<i>M.genitalium</i>	GCU	33,3	48,9	0,507					
	GCC	0,0	7,2	1,000					
	GCA	66,7	39,8	0,168					
	GCG	0,0	4,1	1,000					
<i>B.burgdorferi</i>	GCU	22,2	50,3	0,177					
	GCC	0,0	8,0	1,000					
	GCA	77,8	36,5	0,015					
	GCG	0,0	5,2	1,000					

Eelnevalt on näidatud, et initsiatsioonikoodonile vahetult järgnev koodon võib mõjutada geeni ekspressioonitaset. *S. cerevisiae*'s on suurenenud ekspressioonitasemega seotud UCU

koodon (Miyasaka, 1999). Käesoleva töö koodonkasutuse analüüside tulemustest on samuti näha, et UCU koodoni sagedus on pagaripärmi KEG-ide teises koodoni positsioonis suurenenud võrreldes kõigi geenidega, kuigi GCN sageduste tõus on kõige ulatuslikum (tabel 2). *E. coli*'s on näidatud, et NGG koodonid põhjustavad madala ekspressioonitaseme (Gonzalez de Valdivia ja Isaksson, 2004). Seda peegeldab G kui teise koodoni kolmanda nukleotiidi sageduse tugev langus KEG-ides (joonis 3). Kuna AAA koodon on *E. coli*'s seostatud kõrge ekspressioonitasemega (Stenstrom jt., 2001), siis on antud analüüsist tulenevalt huvitav, et AAA koodoni sagedus ei ole KEG-ide teises koodoni positsioonis tõusnud. A nukleotiidi sagedus teise koodoni esimese nukleotiidina on *E. coli*'s isegi langenud ning langus esineb veel nelja bakteri ja kõigi kolme analüüsitud eukarüooti teise koodoni esimeses või teises nukleotiidi positsioonis (joonis 3).

Üheks võimalikuks põhjuseks võib olla seos peptidüül-tRNA *drop-off*i sagedusega. Koodoneid, mille esimeseks või teiseks nukleotiidiks on A nukleotiid, lugevad peptidüül-tRNA-d on kõige suurema *drop-off*i tõenäosusega (Cruz-Vera jt., 2003). Seetõttu on võimalik, et ORF-ide alguses asuvad A-rikkad koodonid põhjustavad kõrge sagedusega toimuvat peptidüül-tRNA *drop-off*i. Kuna KEG-ide translatsioon moodustab peamise osa raku valgusünteesi aktiivsusest, võidakse A-rikkaid koodoneid ORF-ide alguses vältida seetõttu, et vähendada *drop-off*i produktide hulka rakus. Samuti võivad GCN koodonid olla olulised dipeptidüül-tRNA stabiliseerimiseks ribosoomil.

4. Aminohapete eelistused üherakulistel organismidel

Aminohapete kasutuse võrdlus üherakuliste organismide kõrgelt ekspresseeruvate valkude ja kõigi valkude teises positsioonis näitas, et kõigis uuritud organismides väljaarvatud *M. genitalium* olialaniini sagedus oluliselt tõusnud (tabel 5). Nähtud eelistus oli veelgi tugevam kui GCN koodonite eelistus. Üldine aminohappeline eelistus teises positsioonis oli samuti erinevates organismides sarnane. Lisaksalaniinile olid eelistatud aminohapeteks veel glütsiin ja seriin. Positsioonides 3, 4 ja 5 ei olnud aminohapete eelistus

konserveerunud, kuid neljas genoomis esines nendes positsioonides suurenenud sagedusega positiivselt laetud aminohappeid. Selle põhjuseks võib olla eelnevalt kirjeldatud suurenenud A-rikkus KEG-ide alguses.

Tabel 5. Aminohapete eelistus kõrgelt ekspresseeruvate valkude alguses võrreldes kõigi valkudega (H_0 : aminohapete sagedused ei erine kõrgelt ekspresseeruvates valkudes ja kõigis valkudes).

organism	aminoahappe positsioon											
	2			3			4			5		
	amino-hape	p-väärtus	%KEG/%kõik	amino-hape	p-väärtus	%KEG/%kõik	amino-hape	p-väärtus	%KEG/%kõik	amino-hape	p-väärtus	%KEG/%kõik
<i>E.coli</i>	Ala	3,5E-06	26/9	Lys	0,004	23/11	Ile	7,2E-04	23/9	Lys	3,8E-05	21/7
<i>B.subtilis</i>	Ala	3,6E-06	28/7		-		Gly	0,003	13/3		-	
<i>H.pylori</i>	Ala	1,0E-04	26/8		-			-			-	
<i>H.influenzae</i>	Ala Ser	3,0E-06 0,007	29/9 23/11		-			-		Val	0,008	13/5
<i>M.tuberculosis</i>	Ala	6,4E-09	43/12	Lys	1,3E-04	15/3	Tyr Lys	0,005 0,003	8/2 11/3	Lys Thr	2,5E-04 0,007	13/3 16/6
<i>T.pallidum</i>	Ala	3,7E-05	29/8		-			-			-	
<i>R.prowazekii</i>	Ala Ser	7,6E-04 0,007	18/5 27/11	Thr	0,003	18/5		-			-	
<i>M.genitalium</i>		-			-			-			-	
<i>B.burgdorferi</i>	Ala Gly	3,4E-04 3,4E-04	19/4 15/3		-			-			-	
<i>A.fulgidus</i>	Ala	8,6E-06	26/7		-			-			-	
<i>M.jannaschii</i>	Ala	1,8E-07	23/4		-		Arg	0,009	13/4		-	
<i>P.horikoshii</i>	Ala	0,007	13/4		-		Met	0,004	27/12		-	
<i>S.cerevisiae</i>	Ala Gly	3,5E-07 0,002	28/8 14/5	Arg	1,0E-05	21/6	Val	1,0E-04	18/5	Lys Leu	0,002 0,002	20/9 0/9
<i>S.pombe</i>	Ala Gly	6,4E-05 6,3E-04	28/10 17/5		-			-			-	
<i>P.falciparum</i>	Ala Asn	1,2E-05 0,003	23/5 0/12		-		Gln	0,008	11/3		-	
<i>H.sapiens</i>		-			-		Asn	0,003	1/3	Ile Trp	0,004 0,009	5/3 3/2

Valkude esimesed N-terminaalsed jäägid moduleerivad valkude stabiilsust ning määravad N-terminaalse formüül-metioniini (või eukarüootide puhul metioniini) lõikamise. Seetõttu võib nähtud nukleotiidide ja koodonite eelistuse üheks põhjuseks olla ka vastavate lõikamist soodustavate aminohapete eelistus. Nii bakterites kui eukarüootides on aminohapped, mille eest (formüül-)metioniin ära lõigatakse, sarnased (Ala, Gly, Pro, Ser,

Thr, Val) (Tsunasawa jt., 1985; Ben-Bassat jt., 1987; Miller jt., 1987; Moerschell jt., 1990; Chen jt., 2002). Vastavalt N-terminaalsele reeglile on ka kõik need kuus aminohapet valku stabiliseerivad aminohapped nii bakterites kui *S.cerevisiae*'s (Varshavsky, 1996). Võrreldes kõigi nende aminohapete esinemist kõrgelt ekspresseeruvate valkude ja kõigi valkude alguses selgus, et teises positsioonis N-terminaalset lõikamist soodustavaid ja stabiliseerivaid aminohappeid sisaldavaid valke kodeerivad geenid on KEG-ide hulgas oluliselt rohkem esindatud kui kõigi geenide hulgas (tabel 6). See tulemus sobib kokku ka suure arvu *housekeeping* geenide esinemisega KEG andmestikes.

Tabel 6. Teise aminohappena Ala, Gly, Pro, Ser, Thr ja Val-i sisaldavate valkude osakaal kõrgelt ekspresseeruvates valkudes ja kõigis valkudes (H_0 : selle aminohapete grupi sagedused ei erine kõrgelt ekspresseeruvates valkudes ja kõigis valkudes).

organism	Ala, Ser, Thr, Gly, Val ja Pro protsent 2.positsioonis		p-väärtus
	KEG	kõik	
<i>E.coli</i>	60	40	4,8E-04
<i>B.subtilis</i>	58	32	8,9E-05
<i>H.pylori</i>	56	28	4,1E-05
<i>H.influenzae</i>	64	36	3,2E-06
<i>M.tuberculosis</i>	80	63	0,005
<i>T.pallidum</i>	52	38	0,068
<i>R.prowazekii</i>	71	31	2,8E-08
<i>M.genitalium</i>	44	29	0,040
<i>B.burgdorferi</i>	58	24	1,3E-06
<i>A.fulgidus</i>	55	30	2,0E-04
<i>M.jannaschii</i>	54	26	2,0E-05
<i>P.horikoshii</i>	48	31	0,015
<i>S.cerevisiae</i>	85	54	1,0E-08
<i>S.pombe</i>	90	52	6,6E-10
<i>P.falciparum</i>	79	28	1,5E-12
<i>H.sapiens</i>	60	57	0,039

Kui nähtud järjestuse selektsioon põhineb tõepoolest lõikamist soodustavate ja stabiliseerivate aminohapete valikul, siis on sellegipoolest raske põhjendada, miks on nende kuue sarnaste omadustega aminohappe hulgast välja valitud justalaniin. On võimalik, et teised aminohapped ei ole vastavates mehhanismides nii efektiivsed kuialaniin. Samuti on võimalik, etalaniin aitab tekkival peptiidil siseneda ribosomaalsesse tunnelisse (Tenson ja Ehrenberg, 2002). Ka nähtud positiivselt laetud aminohapete eelistus mitmetes organismides võib olla seotud peptiidi ja ribosoomi interaktsiooniga.

5. Järjestuse analüüs inimese kõrge ekspressioonitasemega ORF-ide alguses

Hulkraksete organismide puhul muudab kõrge ekspressioonitasemega geenide andmestiku koostamise keerukaks ekspressiooni ning seega tõenäoliselt ka koodonkasutuse koespetsiifilisus. Lisaks tuleb arvestada, et käesolevas töös kasutatud analüüside spetsiifika nõuab täpset geeniannotatsiooni ning seetõttu kuuluvad inimese puhul analüüsitud geenide hulka ainult korrektse start- ja stoppkoodoniga kodeerivad järjestused.

Kuigi inimeses ei esinenud G₄ ja C₅ nukleotiidide ülesindatust KEG-ides (tabel 7), olid need nukleotiidid siiski vastavates positsioonides KEG-ides kõige sagedasemad (G₄ 48,6 % ja C₅ 44,0%).

Tabel 7. Nukleotiidide sagedused inimese KEG-ides ja kõigis geenides koodonites 2-5.

nukleotiidi positsioon	U			C			A			G		
	p-väärtus	%KEG	%kõik	p-väärtus	%KEG	%kõik	p-väärtus	%KEG	%kõik	p-väärtus	%KEG	%kõik
4	0,771	13,7	13,4	0,645	15,7	15,2	0,629	21,9	21,3	0,372	48,6	50,1
5	0,810	13,4	13,8	0,060	44,0	40,9	0,028	23,7	27,0	0,669	18,9	18,3
6	0,712	20,8	20,3	0,941	26,7	26,8	0,404	13,9	15,0	0,658	38,6	37,9
7	0,067	14,3	16,5	0,142	28,2	26,0	0,722	23,0	22,5	0,835	34,6	34,9
8	0,112	22,8	20,6	0,057	29,6	32,6	0,344	27,0	25,6	0,686	20,6	21,2
9	0,402	15,9	17,0	0,001	37,8	32,4	0,080	14,2	16,3	0,174	32,1	34,3
10	0,221	18,8	17,2	0,658	27,3	28,0	0,907	23,8	23,7	0,544	30,2	31,1
11	0,001	29,3	24,5	0,008	25,4	29,5	0,849	24,7	25,0	0,745	20,5	21,0
12	0,661	16,6	17,2	0,392	32,5	31,2	0,409	16,5	17,6	0,807	34,4	34,0
13	0,821	16,0	15,7	0,103	23,4	25,8	0,502	27,5	24,8	0,753	33,1	33,7
14	0,005	28,7	24,6	0,302	23,5	25,0	0,520	24,8	25,8	0,282	23,0	24,6
15	7,6E-06	11,1	16,3	0,021	33,6	30,1	0,470	15,1	16,0	0,102	40,2	37,5

Lisaks olid G₄ ja C₅ ka kõige levinuimad nukleotiidid kõigi geenide grupis (G₄ 50,1% ja C₅ 40,9%). Võrreldes genoomi keskmise G ja C nukleotiidide sagedusega (G 26,4% ja C 25,9%) on nii KEG-ides kui ka kõigis geenides selline eelistus teiste nukleotiidide ees statistiliselt väga oluline ($p \ll 0,001$). Tulemus langeb kokku ka nn. Kozaki konsensussega ((G/A)CCAUGG) (Kozak, 1991). Inimese KEG-ide teises ja kolmandas koodoni positsioonis ei olnud ühegi koodoni sagedus võrreldes kõigi geenidega oluliselt suurenenud või vähenenud (tabel 2).

Lisaks analüüsisime ka inimese kõrgelt ekspresseeruvate valkude aminohapete kasutust (tabel 5). Teises aminohappe positsioonis esinevate aminohapete üle- ja alaeesindatuse statistilised olulisused olid madalamad kui üherakuliste organismide puhul ($p > 0,01$). Kuigi erinevalt üherakulistest organismidest ei esinenud inimese kõrgelt ekspresseeruvate valkude teises aminohappe positsioonis alaniini ülesindatust, oli alaniin selles positsioonis siiski levinuim aminohape (23,7%) ning tema sagedus oli sarnane üherakuliste organismide kõrgelt ekspresseeruvate valkude teise positsiooni alaniini sagedustega. Erinevalt üherakulistest organismidest on aga inimeses alaniin ülekaalukalt eelistatuim aminohape ka kõigi valkude teise aminohappena (23,1%). Erinevatest alaniini koodonitest on inimese KEG-ide teises positsioonis võrreldes kogu KEG-idega tugevalt ülesindatud GCG koodon (tabel 4).

Mitmerakulistes organismides võib koodonkasutus olla erinevates kudedes erinev. Seetõttu võib põhjus, miks me ei näinud inimese kõrgelt ekspresseeruvates valkudes alaniini ülesindatust, ühelt poolt olla seotud inimese KEG andmestiku valikuga, kuna andmestik sisaldab ka koetsütofilisi geene. Lisaks jäi kindlasti inimese KEG-ide hulgest kõrge ekspressioonitasemega geene välja ORF-ide korrektse alguskoordinaadi puudumise tõttu. Samas võib alaniini sageduse tõusu puhul olla tegu hoopis laiemas ja isegi tugevama efektiga kui üherakulistes organismides, hõlmates inimese puhul kõiki organismi geene.

6. Algandmestiku valiku kontroll

Käesoleva töö jaoks organisme valides püüdsime luua andmestikku, mis sisaldaks võimalikult erinevate omadustega genome (genoomi suurus, GC-sisaldus, elukeskkond). Kontrolliks, et nähtudalaniini ülesindatus pole siiski vaid antud andmestiku omapära, analüüsisime veel lisaks kõigi sekveneeritud genomide hulgast juhuslikult valitud kolme bakteri (*Mycobacterium avium paratuberculosis*, *Synechococcus elongatus*, *Rickettsia conorii*), ühte arhea (*Haloarcula marismortui*) ja ühte üherakulise eukarüooti (*Cryptococcus neoformans*) genoomi. Ka nende genomide puhul on KEG andmestikud koostatud ortoloogide meetodi põhjal (lisa IV,V). Neljas viiest lisaks analüüsitud genomist esines statistiliselt olulinealaniini ülesindatus kõrgelt ekspresseeruvate valkude teises aminohappe positsioonis (tabel 8). Ka *R. conorii* KEG-ide poolt kodeeritud valkudes olialaniini sagedus 2,3 korda tõusnud võrreldes kõigi valkudega, kuigi erinevuse statistiline olulisus oli madal ($p = 0,043$). Need tulemused viitavad, et nähtud tendents esineb enamuses üherakulistest organismidest.

Tabel 8. Aminohapete eelistused juhuslikult valitud organismide kõrgelt ekspresseeruvate valkude alguses võrreldes kõigi valkudega (H_0 : aminohapete sagedused ei erine kõrgelt ekspresseeruvates valkudes ja kõigis valkudes).

organism	aminohappe positsioon											
	2			3			4			5		
	amino-hape	p-väärtus	%KEG/%kõik	amino-hape	p-väärtus	%KEG/%kõik	amino-hape	p-väärtus	%KEG/%kõik	amino-hape	p-väärtus	%KEG/%kõik
<i>S.elongatus</i>	Ala	2,4E-04	34/14	Lys	1,5E-05	22/5	Lys	3,9E-04	14/3	-	-	-
<i>M.avium paratuberculosis</i>	Ala	8,6E-10	46/14	Lys	1,3E-04	14/3	Lys	4,3E-05	14/3	Lys	2,3E-04	13/3
<i>R.conorii</i>	-	-	-	Thr	9,0E-04	17/4	-	-	-	-	-	-
<i>H.marismortui</i>	Ala	6,8E-06	31/11	-	-	-	Glu	0,004	18/7	-	-	-
<i>C.neoformans</i>	Gly	0,001	14/4	Lys	0,002	14/4	Val	0,005	13/4	Lys	5,2E-04	17/5
	Ala	0,003	28/14	-	-	-	Lys	0,009	13/5	-	-	-

7. Ortoloogidel põhineva kõrgelt ekspresseeruvate geenide leidmise meetodi kontroll

Kuigi ortoloogide meetodit on organismi kõrgelt ekspresseeruvate geenide määramiseks väga edukalt kasutatud ka eelnevalt (McVean and Hurst 2000, Perriere and Thioulouse 2002), teostasime siiski lisaanalüüse veendumaks, et kasutatud meetod on tõepoolest sobiv. Organismi sünonüümset koodonkasutust saab analüüsida korrespondentsanalüüsiga, mis paigutab geenid 61-dimensionaalsesse ruumi. Seejuures klasterduvad sarnase sünonüümse koodonkasutusega geenid kokku. Teostasime vastava analüüsi kõigi valitud üherakuliste organismide peal. Selgus, et kõigi üherakuliste organismide puhul grupeerusid ortoloogide meetodil leitud eeldatavad KEG-id kokku ribosomaalsete valkude geenidega, mis on teatavasti kõrge ekspressioonitasemega (Dong jt., 1996) (näiteid lisas VI). Seega on leitud ortoloogide puhul tegemist tõesti kõrge ekspressioonitasemega geenidega. Lisaks nägime, et organismides, kus KEG-id on sünonüümse koodonkasutuse alusel teistest geenidest selgelt eraldunud, leidub veel geene, millel on ortoloogidega sarnane koodonkasutus (lisa VI). Juhul, kui nähtudalaniini ülesindatus on tõepoolest omane kõrge ekspressioonitasemega geenidele, peaks muster esinema ka nendes lisageenides. Selle kontrollimiseks analüüsisime aminohapete kasutust *B. subtilis*'e, *H. influenzae*, *M. jannaschii* ja *S. pombe* vastavates kõrgelt ekspresseeruvates geenides (lisa VII). Tulemustest selgus, et *B. subtilis*'e ja *S. pombe* puhul onalaniini ülesindatuse p-väärtused veelgi vähenenud (tabel 9). *M. jannaschii*'s p-väärtus tõusis, kuid oli siiski statistiliselt oluline. *H. influenzae*'s suurenes p-väärtus 0,03-ni (alaniini ülesindatus ei ole statistiliselt oluline). Seega esineb ka lisa-analüüsides kasutatud kõrge ekspressioonitasemega geenidesalaniini ülesindatus enamuses organismides.

Tabel 9. Aminohapete eelistused lisaks analüüsitud kõrgelt ekspresseeruvates valkudes võrreldes kõigi valkudega.

organism	aminohappe positsioon											
	2			3			4			5		
	amino- hape	p-väärtus	%KEG/ %kõik	amino- hape	p-väärtus	%KEG/ %kõik	amino- hape	p-väärtus	%KEG/ %kõik	amino- hape	p-väärtus	%KEG/ %kõik
<i>B.subtilis</i>	Ala	2,6E-08	24/7		-			-			-	
<i>H.influenzae</i>		-			-			-			-	
<i>M.jannaschii</i>	Ala	2,0E-05	14/4		-			-			-	
<i>S.pombe</i>	Ala	4,8E-06	25/10	Ala	0,008	12/5		-		Lys	0,004	17/8
	Asn	0,002	0/6									

KOKKUVÕTE

Käesoleva töö eesmärgiks oli uurida, kas kõrge ekspressioonitasemega geenidel esineb vahetult startkoodoni järel asuvas järjestuses iseloomulikke mustreid. Selleks võrreldi detailsetes nukleotiidi, koodonite ja aminohapete kasutuse analüüsides erinevate organismide kõrge ekspressioonitasemega geenide gruppe organismi kõigi geenide grupiga. Töös analüüsiti kokku 9 bakteri, 3 arhea, 3 üherakulise eukarüooti ja inimese kõrge ekspressioonitasemega geenide alguste järjestusi. Eksperimentaalsete geeniekspressiooniandmete puudumise tõttu kasutati erinevate organismide kõrge ekspressioonitasemega geenide grupi väljaselgitamiseks üherakuliste organismide puhul ortoloogidel põhinevat lähenemist. Kontrollkatsed vastava meetodi kasutamise õigsuse kinnitamiseks näitasid, et leitud ja analüüsides kasutatud *E. coli* ja *S.cerevisiae* kõrge ekspressioonitasemega geenide ortoloogid on tõepoolest korrektne andmestik. Esiteks langesid leitud ortoloogid koodonkasutuse alusel ühte gruppi ribosomaalsete geenidega. Teiseks, lisaks analüüsitud kõrge ekspressioonitasemega geenide grupid andsid esialgsete gruppidega väga sarnased analüüsides tulemused.

Töö tulemusena leiti, et erinevate organismide KEG-id sisaldavad suurenenud sagedusega C nukleotiidi viiendas ja G nukleotiidi neljandas ORF-ide positsioonis võrreldes kõigi geenidega. Tulemus langeb kokku osaga klassikalisest Kozaki konsensusest (GCCRCCAAUGG) (Kozak, 1997). Huvitav on seejuures asjaolu, et see konsensusosa esineb ka bakterites ja arheates.

G₄ ja C₅ nukleotiidide ülesindatus korreleerub KEG-ide teise positsiooni koodonkasutusega, kus on eelistatud GCN koodonid. KEG-ide poolt kodeeritud valkude aminohapete kasutuse muster oli aga veelgi tugevam: üherakulistes organismides esines tugevalaniini ülesindatus kõrgelt ekspresseeruvate valkude teises aminohappe positsioonis. Alaniini eelistus esines kõigis uuritud üherakulistes genoomides v.a *M. genitalium*, mis viitab bakterite, arheate ja üherakuliste eukarüootide kõrge ekspressioonitasemega geenide universaalsele omadusele. Inimese kõrgelt

ekspresseeruvates valkudes ei esinenud teises positsioonis ühegi aminohappe olulist ülesindatust. Siiski on inimese nii kõrgelt ekspresseeruvates kui ka kõigis valkudes levinuimaks teiseks aminohappeks samutialaniin, mis võib viidata üherakulistes organismides leiduva mustri sarnasele, kuid inimeses veel laiemalt levinud mustrile.

SUMMARY

The aim of the current work was to study whether the highly expressed genes have specific patterns in the sequences immediately following the start codon. Therefore we compared the frequencies of nucleotides, codons and amino acids in highly expressed genes with the corresponding frequencies in all genes of studied organisms.

We analysed the beginnings of highly expressed genes in 9 bacteria, 3 archaea, 3 unicellular eukaryotes and human. Due to the lack of experimental gene expression data we used the method based on orthologues to compile the groups of highly expressed genes of different organisms. In additional analysis we ensured that the orthologues of the highly expressed genes of *E. coli* and *S. cerevisiae* are indeed appropriate dataset. Firstly, the orthologues we found clustered together with the genes for ribosomal proteins. Secondly, the analysis with additional highly expressed genes groups gave very similar results to the initial groups.

Our work revealed that highly expressed genes, when compared to the all genes dataset contain increased frequency of G nucleotide in the 4th position and C in the 5th position of the ORFs. The result is consistent with the part of Kozak consensus (GCCRCCAAUGG) (Kozak, 1997). Interestingly, this part of consensus is also present in bacteria and archaea.

The overrepresentation of G₄ and C₅ is correlated with the codon usage in the second position of highly expressed genes where GCN codons are preferred. The amino acids usage pattern in proteins coded by highly expressed genes was even stronger: in unicellular organisms alanine was strongly overrepresented in the second amino acid position of highly expressed proteins. The preference for alanine was present in all studied unicellular genomes, except *M.genitalium*. This refers to the universal characteristic of highly expressed genes in bacteria, archaea and unicellular eukaryotes. None of the amino acids was significantly preferred in the second position of human highly expressed proteins. Still, the most frequent second amino acid in human highly expressed proteins and all proteins

was alanine. This could be explained by the similar but more extended pattern in humans than in unicellular organisms.

TÄNUAVALDUSED

Suured tänud juhendajatele prof. Maida Remmile ja dr. Tanel Tensonile heade nõuannete ja kannatlikkuse eest. Samuti Tõnu Mölsile abi eest statistiliste meetodite valikul. Tänan ka kallist elukaaslast ja sõpru toetuse eest.

VIITED

- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. ja Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.
- Andersson, S.G. ja Kurland, C.G. (1990). Codon preferences in free-living microorganisms. *Microbiol Rev.* 54: 198-210.
- Ben-Bassat, A., Bauer, K., Chang, S.Y., Myambo, K., Boosman, A. ja Chang, S. (1987). Processing of the initiation methionine from proteins: properties of the *Escherichia coli* methionine aminopeptidase and its gene structure. *J Bacteriol.* 169: 751-757.
- Benelli, D., Maone, E. ja Londei, P. (2003). Two different mechanisms for ribosome/mRNA interaction in archaeal translation initiation. *Mol Microbiol.* 50: 635-643.
- Bennetzen, J.L. ja Hall, B.D. (1982). Codon selection in yeast. *J Biol Chem.* 257: 3026-3031.
- Blattner, F.R., Plunkett, G., 3rd, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F. jt. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science.* 277: 1453-1474.
- Boni, I.V., Isaeva, D.M., Musychenko, M.L. ja Tzareva, N.V. (1991). Ribosome-messenger recognition: mRNA target sites for ribosomal protein S1. *Nucleic Acids Res.* 19: 155-162.
- Cavener, D.R. ja Ray, S.C. (1991). Eukaryotic start and stop translation sites. *Nucleic Acids Res.* 19: 3185-3192.
- Chen, G.F. ja Inouye, M. (1990). Suppression of the negative effect of minor arginine codons on gene expression; preferential usage of minor codons within the first 25 codons of the *Escherichia coli* genes. *Nucleic Acids Res.* 18: 1465-1473.
- Chen, H., Bjerknes, M., Kumar, R. ja Jay, E. (1994). Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of *Escherichia coli* mRNAs. *Nucleic Acids Res.* 22: 4953-4957.

Chen, S., Vetro, J.A. ja Chang, Y.H. (2002). The specificity in vivo of two distinct methionine aminopeptidases in *Saccharomyces cerevisiae*. *Arch Biochem Biophys.* 398: 87-93.

Cigan, A.M. ja Donahue, T.F. (1987). Sequence and structural features associated with translational initiator regions in yeast--a review. *Gene.* 59: 1-18.

Cigan, A.M., Feng, L. ja Donahue, T.F. (1988) tRNAⁱ(met) functions in directing the scanning ribosome to the start site of translation. *Science.* 93-97.

Coghlan, A. ja Wolfe, K.H. (2000). Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. *Yeast.* 16: 1131-1145.

Cruz-Vera, L.R., Hernandez-Ramon, E., Perez-Zamorano, B. ja Guarneros, G. (2003). The rate of peptidyl-tRNA dissociation from the ribosome during minigene expression depends on the nature of the last decoding interaction. *J Biol Chem.* 278: 26065-26070.

Curran, J. ja Kolakofsky, D. (1988). Scanning independent ribosomal initiation of the Sendai virus X protein. *Embo J.* 7: 2869-2874.

Dabrowski, C. ja Alwine, J.C. (1988). Translational control of synthesis of simian virus 40 late proteins from polycistronic 19S late mRNA. *J Virol.* 62: 3182-3192.

Dalboge, H., Bayne, S. ja Pedersen, J. (1990). In vivo processing of N-terminal methionine in *E. coli*. *FEBS Lett.* 266: 1-3.

de Smit, M.H. ja van Duin, J. (1994). Translational initiation on structured messengers. Another role for the Shine-Dalgarno interaction. *J Mol Biol.* 235: 173-184.

Dincbas, V., Heurgue-Hamard, V., Buckingham, R.H., Karimi, R. ja Ehrenberg, M. (1999). Shutdown in protein synthesis due to the expression of mini-genes in bacteria. *J Mol Biol.* 291: 745-759.

Dong, H., Nilsson, L. ja Kurland, C.G. (1996). Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *J Mol Biol.* 260: 649-663.

dos Reis, M., Wernisch, L. ja Savva, R. (2003). Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome. *Nucleic Acids Res.* 31: 6976-6985.

- Esposito, D., Fey, J.P., Eberhard, S., Hicks, A.J. ja Stern, D.B. (2003). In vivo evidence for the prokaryotic model of extended codon-anticodon interaction in translation initiation. *Embo J.* 22: 651-656.
- Eyre-Walker, A. ja Bulmer, M. (1993). Reduced synonymous substitution rate at the start of enterobacterial genes. *Nucleic Acids Res.* 21: 4599-4603.
- Farwell, M.A., Roberts, M.W. ja Rabinowitz, J.C. (1992). The effect of ribosomal protein S1 from *Escherichia coli* and *Micrococcus luteus* on protein synthesis in vitro by *E. coli* and *Bacillus subtilis*. *Mol Microbiol.* 6: 3375-3383.
- Firpo, M.A. ja Dahlberg, A.E. (1998). The importance of base pairing in the penultimate stem of *Escherichia coli* 16S rRNA for ribosomal subunit association. *Nucleic Acids Res.* 26: 2156-2160.
- Fuglsang, A. (2004). Bioinformatic analysis of the link between gene composition and expressivity in *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. *Antonie Van Leeuwenhoek.* 86: 135-147.
- Futcher, B., Latter, G.I., Monardo, P., McLaughlin, C.S. ja Garrels, J.I. (1999). A sampling of the yeast proteome. *Mol Cell Biol.* 19: 7357-7368.
- Gillman, E.C., Slusher, L.B., Martin, N.C. ja Hopper, A.K. (1991). MOD5 translation initiation sites determine N⁶-isopentenyladenosine modification of mitochondrial and cytoplasmic tRNA. *Mol Cell Biol.* 11: 2382-2390.
- Gonzalez de Valdivia, E.I. ja Isaksson, L.A. (2004). A codon window in mRNA downstream of the initiation codon where NGG codons give strongly reduced gene expression in *Escherichia coli*. *Nucleic Acids Res.* 32: 5198-5205.
- Gualerzi, C.O. ja Pon, C.L. (1990). Initiation of mRNA translation in prokaryotes. *Biochemistry.* 29: 5881-5889.
- Gutell, R.R., Weiser, B., Woese, C.R. ja Noller, H.F. (1985). Comparative anatomy of 16S-like ribosomal RNA. *Prog Nucleic Acid Res Mol Biol.* 32: 155-216.
- Gutierrez, G., Marquez, L. ja Marin, A. (1996). Preference for guanosine at first codon position in highly expressed *Escherichia coli* genes. A relationship with translational efficiency. *Nucleic Acids Res.* 24: 2525-2527.

- Hamilton, R., Watanabe, C.K. ja de Boer, H.A. (1987). Compilation and comparison of the sequence context around the AUG startcodons in *Saccharomyces cerevisiae* mRNAs. *Nucleic Acids Res.* 15: 3581-3593.
- Hernandez-Sanchez, J., Valadez, J.G., Herrera, J.V., Ontiveros, C. ja Guarneros, G. (1998). lambda bar minigene-mediated inhibition of protein synthesis involves accumulation of peptidyl-tRNA and starvation for tRNA. *Embo J.* 17: 3758-3765.
- Heurgue-Hamard, V., Dincbas, V., Buckingham, R.H. ja Ehrenberg, M. (2000). Origins of minigene-dependent growth inhibition in bacterial cells. *Embo J.* 19: 2701-2709.
- Hirel, P.H., Schmitter, M.J., Dessen, P., Fayat, G. ja Blanquet, S. (1989). Extent of N-terminal methionine excision from *Escherichia coli* proteins is governed by the side-chain length of the penultimate amino acid. *Proc Natl Acad Sci U S A.* 86: 8247-8251.
- Ikemura, T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J Mol Biol.* 151: 389-409.
- Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol.* 2: 13-34.
- Jansen, R., Bussemaker, H.J. ja Gerstein, M. (2003). Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. *Nucleic Acids Res.* 31: 2242-2251.
- Jia, M. ja Li, Y. (2005). The relationship among gene expression, folding free energy and codon usage bias in *Escherichia coli*. *FEBS Lett.* 579: 5333-5337.
- Karlin, S. ja Mrazek, J. (2000). Predicted highly expressed genes of diverse prokaryotic genomes. *J Bacteriol.* 182: 5238-5250.
- Karlin, S., Mrazek, J., Ma, J. ja Brocchieri, L. (2005). Predicted highly expressed genes in archaeal genomes. *Proc Natl Acad Sci U S A.* 102: 7303-7308.
- Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J. jt. (2003). The UCSC Genome Browser Database. *Nucleic Acids Res.* 31: 51-54.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. ja Haussler, D. (2002). The human genome browser at UCSC. *Genome Res.* 12: 996-1006.

Komarova, A.V., Tchufistova, L.S., Dreyfus, M. ja Boni, I.V. (2005). AU-rich sequences within 5' untranslated leaders enhance translation and stabilize mRNA in *Escherichia coli*. *J Bacteriol.* 187: 1344-1349.

Komarova, A.V., Tchufistova, L.S., Supina, E.V. ja Boni, I.V. (2002). Protein S1 counteracts the inhibitory effect of the extended Shine-Dalgarno sequence on translation. *Rna.* 8: 1137-1147.

Kozak, M. (1983). Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. *Microbiol Rev.* 47: 1-45.

Kozak, M. (1987). An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.* 15: 8125-8148.

Kozak, M. (1989). Circumstances and mechanisms of inhibition of translation by secondary structure in eucaryotic mRNAs. *Mol Cell Biol.* 9: 5134-5142.

Kozak, M. (1991). Structural features in eukaryotic mRNAs that modulate the initiation of translation. *J Biol Chem.* 266: 19867-19870.

Kozak, M. (1995). Adherence to the first-AUG rule when a second AUG codon follows closely upon the first. *Proc Natl Acad Sci U S A.* 92: 7134.

Kozak, M. (1997). Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6. *Embo J.* 16: 2482-2492.

Kozak, M. (1999). Initiation of translation in prokaryotes and eukaryotes. *Gene.* 234: 187-208.

Kutay, U., Ahnert-Hilger, G., Hartmann, E., Wiedenmann, B. ja Rapoport, T.A. (1995). Transport route for synaptobrevin via a novel pathway of insertion into the endoplasmic reticulum membrane. *Embo J.* 14: 217-223.

Le, S.Y. ja Maizel, J.V., Jr. (1997). A common RNA structural motif involved in the internal initiation of translation of cellular mRNAs. *Nucleic Acids Res.* 25: 362-369.

Martinez-Salas, E. (1999). Internal ribosome entry site biology and its use in expression vectors. *Curr Opin Biotechnol.* 10: 458-464.

Mazel, D., Pochet, S. ja Marliere, P. (1994). Genetic characterization of polypeptide deformylase, a distinctive enzyme of eubacterial translation. *Embo J.* 13: 914-923.

Meinzel, T., Mechulam, Y. ja Blanquet, S. (1993). Methionine as translation start signal: a review of the enzymes of the pathway in *Escherichia coli*. *Biochimie.* 75: 1061-1075.

Menez, J., Heurgue-Hamard, V. ja Buckingham, R.H. (2000). Sequestration of specific tRNA species cognate to the last sense codon of an overproduced gratuitous protein. *Nucleic Acids Res.* 28: 4725-4732.

Miller, C.G., Strauch, K.L., Kukral, A.M., Miller, J.L., Wingfield, P.T., Mazzei, G.J., Werlen, R.C., Graber, P. ja Movva, N.R. (1987). N-terminal methionine-specific peptidase in *Salmonella typhimurium*. *Proc Natl Acad Sci U S A.* 84: 2718-2722.

Miyasaka, H. (1999). The positive relationship between codon usage bias and translation initiation AUG context in *Saccharomyces cerevisiae*. *Yeast.* 15: 633-637.

Moerschell, R.P., Hosokawa, Y., Tsunasawa, S. ja Sherman, F. (1990). The specificities of yeast methionine aminopeptidase and acetylation of amino-terminal methionine in vivo. Processing of altered iso-1-cytochromes c created by oligonucleotide transformation. *J Biol Chem.* 265: 19638-19643.

Niimura, Y., Terabe, M., Gojobori, T. ja Miura, K. (2003). Comparative analysis of the base biases at the gene terminal portions in seven eukaryote genomes. *Nucleic Acids Res.* 31: 5195-5201.

O'Connor, M., Asai, T., Squires, C.L. ja Dahlberg, A.E. (1999). Enhancement of translation by the downstream box does not involve base pairing of mRNA with the penultimate stem sequence of 16S rRNA. *Proc Natl Acad Sci U S A.* 96: 8973-8978.

Ohno, H., Sakai, H., Washio, T. ja Tomita, M. (2001). Preferential usage of some minor codons in bacteria. *Gene.* 276: 107-115.

Remm, M., Storm, C.E. ja Sonnhammer, E.L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol.* 314: 1041-1052.

Roberts, M.W. ja Rabinowitz, J.C. (1989). The effect of *Escherichia coli* ribosomal protein S1 on the translational specificity of bacterial ribosomes. *J Biol Chem.* 264: 2228-2235.

Rocha, E.P., Danchin, A. ja Viari, A. (1999). Translation in *Bacillus subtilis*: roles and trends of initiation and termination, insights from a genome analysis. *Nucleic Acids Res.* 27: 3567-3576.

Sakai, H., Imamura, C., Osada, Y., Saito, R., Washio, T. ja Tomita, M. (2001). Correlation between Shine--Dalgarno sequence conservation and codon usage of bacterial genes. *J Mol Evol.* 52: 164-170.

Schatz, P.J. ja Beckwith, J. (1990). Genetic analysis of protein export in *Escherichia coli*. *Annu Rev Genet.* 24: 215-248.

Schmitt, E., Guillon, J.M., Meinnel, T., Mechulam, Y., Dardel, F. ja Blanquet, S. (1996). Molecular recognition governing the initiation of translation in *Escherichia coli*. A review. *Biochimie.* 78: 543-554.

Sedman, S.A., Gelembiuk, G.W. ja Mertz, J.E. (1990). Translation initiation at a downstream AUG occurs with increased efficiency when the upstream AUG is located very close to the 5' cap. *J Virol.* 64: 453-457.

Sharp, P.M. ja Li, W.H. (1987). The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15: 1281-1295.

Sherman, F., Stewart, J.W. ja Tsunasawa, S. (1985). Methionine or not methionine at the beginning of a protein. *Bioessays.* 3: 27-31.

Shine, J. ja Dalgarno, L. (1974). The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci U S A.* 71: 1342-1346.

Slusher, L.B., Gillman, E.C., Martin, N.C. ja Hopper, A.K. (1991). mRNA leader length and initiation codon context determine alternative AUG selection for the yeast gene MOD5. *Proc Natl Acad Sci U S A.* 88: 9789-9793.

Solbiati, J., Chapman-Smith, A., Miller, J.L., Miller, C.G. ja Cronan, J.E., Jr. (1999). Processing of the N termini of nascent polypeptide chains requires deformylation prior to methionine removal. *J Mol Biol.* 290: 607-614.

Steitz, J.A. ja Jakes, K. (1975). How ribosomes select initiator regions in mRNA: base pair formation between the 3' terminus of 16S rRNA and the mRNA during initiation of protein synthesis in *Escherichia coli*. *Proc Natl Acad Sci U S A.* 72: 4734-4738.

Stenstrom, C.M., Holmgren, E. ja Isaksson, L.A. (2001). Cooperative effects by the initiation codon and its flanking regions on translation initiation. *Gene.* 273: 259-265.

Stenstrom, C.M. ja Isaksson, L.A. (2002). Influences on translation initiation and early elongation by the messenger RNA region flanking the initiation codon at the 3' side. *Gene.* 288: 1-8.

Stenstrom, C.M., Jin, H., Major, L.L., Tate, W.P. ja Isaksson, L.A. (2001). Codon bias at the 3'-side of the initiation codon is correlated with translation initiation efficiency in *Escherichia coli*. *Gene.* 263: 273-284.

Stormo, G.D., Schneider, T.D. ja Gold, L.M. (1982). Characterization of translational initiation sites in *E. coli*. *Nucleic Acids Res.* 10: 2971-2996.

Supek, F. ja Vlahovicek, K. (2005). Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. *BMC Bioinformatics.* 6: 182.

Tenson, T. ja Ehrenberg, M. (2002). Regulatory nascent peptides in the ribosomal tunnel. *Cell.* 108: 591-594.

Tenson, T., Herrera, J.V., Kloss, P., Guarneros, G. ja Mankin, A.S. (1999). Inhibition of translation and cell growth by minigene expression. *J Bacteriol.* 181: 1617-1622.

Torarinsson, E., Klenk, H.P. ja Garrett, R.A. (2005). Divergent transcriptional and translational signals in Archaea. *Environ Microbiol.* 7: 47-54.

Tsunasawa, S., Izu, Y., Miyagi, M. ja Kato, I. (1997). Methionine aminopeptidase from the hyperthermophilic Archaeon *Pyrococcus furiosus*: molecular cloning and overexpression in *Escherichia coli* of the gene, and characteristics of the enzyme. *J Biochem (Tokyo).* 122: 843-850.

Tsunasawa, S., Stewart, J.W. ja Sherman, F. (1985). Amino-terminal processing of mutant forms of yeast iso-1-cytochrome c. The specificities of methionine aminopeptidase and acetyltransferase. *J Biol Chem.* 260: 5382-5391.

Tzareva, N.V., Makhno, V.I. ja Boni, I.V. (1994). Ribosome-messenger recognition in the absence of the Shine-Dalgarno interactions. *FEBS Lett.* 337: 189-194.

Wang, H.C. ja Hickey, D.A. (2002). Evidence for strong selective constraint acting on the nucleotide composition of 16S ribosomal RNA genes. *Nucleic Acids Res.* 30: 2501-2507.

Varshavsky, A. (1996). The N-end rule: functions, mysteries, uses. *Proc Natl Acad Sci U S A.* 93: 12142-12149.

Velculescu, V.E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M.A., Bassett, D.E., Jr., Hieter, P., Vogelstein, B. ja Kinzler, K.W. (1997). Characterization of the yeast transcriptome. *Cell.* 88: 243-251.

Velculescu, V.E., Zhang, L., Vogelstein, B. ja Kinzler, K.W. (1995). Serial analysis of gene expression. *Science.* 270: 484-487.

Wheeler, D.L., Church, D.M., Federhen, S., Lash, A.E., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E., Tatusova, T.A. jt. (2003). Database resources of the National Center for Biotechnology. *Nucleic Acids Res.* 31: 28-33.

von Heijne, G. (1986). Towards a comparative anatomy of N-terminal topogenic protein sequences. *J Mol Biol.* 189: 239-242.

Wright, F. (1990). The 'effective number of codons' used in a gene. *Gene.* 87: 23-29.

Yoon, H. ja Donahue, T.F. (1992). Control of translation initiation in *Saccharomyces cerevisiae*. *Mol Microbiol.* 6: 1413-1419.

Yueh, A. ja Schneider, R.J. (1996). Selective translation initiation by ribosome jumping in adenovirus-infected and heat-shocked cells. *Genes Dev.* 10: 1557-1567.

LISAD

Lisa I. *E.coli* KEG-ide ortoloogid.

<i>E.coli</i> (80)	<i>A.fulgidus</i> (21)	<i>P.horikoshii</i> (28)	<i>M.jannaschii</i> (23)	<i>B.subtilis</i> (53)	<i>H.influenzae</i> (70)	<i>H.pylori</i> (54)	<i>M.tuberculosis</i> (61)	<i>T.pallidum</i> (48)	<i>R.prowazekii</i> (49)	<i>M.genitalium</i> (45)	<i>B.burgdorferi</i> (48)
lpp											
rplL				+	+	+	+	+	+	+	+
eno	+	+	+	+	+	+	+	+		+	+
gapA	+	+	+	+	+	+	+	+		+	+
ompC					+						
tufA		+			+						
cspA				+							
ahpC	+	+	+	+		+	+	+	+		
tufB		+		+	+	+	+	+	+	+	+
mopA	+	+	+	+	+	+	+	+	+	+	+
ompA					+						
rpsI	+	+	+	+	+	+	+	+	+	+	+
rpsA				+	+	+	+	+	+		+
pflB					+						
rpsB	+	+		+	+	+	+	+	+	+	+
fba					+		+				+
tsf				+	+	+	+	+	+	+	+
rplA	+	+	+	+	+	+	+	+	+	+	+
rplY					+		+		+		
fusA	+	+	+		+	+	+	+	+	+	+
rpmA				+	+	+	+	+	+	+	+
tpiA	+	+	+	+	+	+	+	+		+	+
rpmE				+	+	+	+	+		+	
ompX											
tig				+	+	+	+	+	+	+	+
pgk	+	+	+	+	+	+	+	+		+	+
rpsC	+	+	+	+	+	+	+	+	+	+	+
rpmG					+	+	+	+	+		+
rpsU					+	+		+			+
rpmH					+	+	+	+	+	+	+
rplI				+	+	+	+	+	+	+	+
dnak				+	+	+	+	+	+	+	+
sodA				+	+		+		+		+
rpmI					+	+	+	+		+	+
yaiU											
rplB	+	+	+	+	+	+	+	+	+	+	+
rplC	+	+	+	+	+	+	+	+	+	+	+
yfiD					+						
rpsQ	+	+	+	+	+	+	+	+	+	+	+
pykF		+	+	+			+			+	+

(jätkub)

<i>E.coli</i> (80)	<i>A.fulgidus</i> (21)	<i>P.horikoshii</i> (28)	<i>M.jannaschii</i> (23)	<i>B.subtilis</i> (53)	<i>H.influenzae</i> (70)	<i>H.pylori</i> (54)	<i>M.tuberculosis</i> (61)	<i>T.pallidum</i> (48)	<i>R.prowazekii</i> (49)	<i>M.genitalium</i> (45)	<i>B.burgdorferi</i> (48)
rplK	+	+	+	+	+	+	+	+	+	+	+
rplO				+	+	+	+	+	+	+	+
rplD				+	+	+	+	+	+	+	+
rpoC	+	+	+	+	+	+	+	+	+	+	+
rpsO				+	++	+	+	+	+	+	+
cspC				+					+		
efp				+	+	+	+	+			+
slyD			+		+	+		+			
pal					+	+			+		
trpL											
tktA			+	+	+					+	
rplT				+	+	+	+	+	+	+	+
rpsF					+	+	+		+		
aceE					+		+				
atpA				+	+	+	+		+	+	
ppiB					+		+				
acpP				+	+	+	+	+	+		+
pta				+	+		+		+		
rpsR				+	+	+	+	+	+	+	+
glyA	+	+	+	+	+	+	+	+	+	+	+
hupA				+	+	+	+	+	+		
rplM	+	+	+	+	+	+	+	+	+	+	+
rpsT					+	+	+		+		+
ompF											
pfkA				+	+		+	+		+	+
rpsL	+	+	+	+	+	+	+	+	+	+	+
ackA					+	+	+	+	+	+	+
rpsP				+	+	+	+	+	+	+	+
ppa		+			+	+	+		+	+	
deoD				+	+	+		+		+	
lpdA				+	+		+		+	+	
adhE		+									
rplW		+	+		+	+	+	+	+		+
atpD		+		+	+	+	+		+	+	
adk	+	+		+	+	+	+	+	+	+	+
ybaB				+	+	+	+		+		
bipA				+	+	+	+		+		
uxuA				+	+						
ndk	+	+	+	+	+	+	+	+	+		+
yebC				+		+	+	+	+	+	+

Lisa II. *S.cerevisiae* KEG-ide ortoloogid.

<i>S.cerevisiae</i> (80)	<i>S.pombe</i> (60)	<i>P.falciparum</i> (47)	<i>A.fulgidus</i> (32)	<i>P.horikoshii</i> (24)	<i>M.jannaschii</i> (33)	<i>S.cerevisiae</i> (80)	<i>S.pombe</i> (60)	<i>P.falciparum</i> (47)	<i>A.fulgidus</i> (32)	<i>P.horikoshii</i> (24)	<i>M.jannaschii</i> (33)
TDH3						RPS9B	+	+	+		+
RPL39	+		+		+	RPS17A		+	+		+
TDH2						RPL17A		+	+		+
PDC1	+					RPS14A	+	+	+	+	+
CDC19	+	+				RPP2A	+		+	+	+
ENO2						EFT1	+	+			
TEF2	+	+	+		+	RPL4A	+				
RPL43A	+	+		+	+	SSA2					
ENO1	+	+				ILV5	+		+		+
TEF1	++	+				RPS10B		+			
CCW12						RPS2	+	+	+	+	+
FBA1	+					EFT2	+				
RPL30	+	+	+	+	+	RPS3	+	+			
RPL22A	+	+				RPL36B	+	+			
TDH1	+	+				RPL35A	+	+			+
RPL8B						RPL42B	+		+		+
RPS6B	+	+	+	+	+	RPP0	+	+	+	+	+
RPS12	+					RPS19A		+	+	+	+
RPL8A	+	+				RPS23B	+				
RPS10A	+					RPL15A			+	+	+
RPS5	+	+	+	+	+	RPS26A		+			
RPS6A	+					RPL34B					
RPL1B	+	+				RPL11A	+		+	+	+
RPL3	+	+				YEF3			+		
RPL5	+	+	+	+	+	RPL9B		+	+	+	+
RPS20	+	+	+	+	+	ASC1	+	+			
RPL10	++	+	+	+	+	RPS13	+	+	+	+	+
RPL1A						RPS18A	+	+	+	+	+
SSB1						RPL38	+	+			
RPL32	+	+	+	+	+	RPL2A	+	+			
TP11	+	+				SSB2	+				
RPS24A	+	+	+			RPL9A	+				
PGK1	+	+				RPS1B	+				+
HYP2		+	+	+	+	RPL42A				+	
RPL4B		+	+	+	+	RPS15	+	+	+	+	+
RPS22A	+		+	+	+	RPL12B	+	+			
RPL18A	+	+	+	+	+	CWP2					
GPM1	+	+				RPS16B	+	+			
RPS31	+		+		+	RPL2B					
ADH1	+					EFB1	+	+			

Lisa III. Nukleotiidide kasutuse p-väärtused erinevates organismides positsioonides 4-30 (koodonid 2-10). (ecoli – *Escherichia coli*, bsubt – *Bacillus subtilis*, hpylo – *Helicobacter pylori*, hinfl – *Haemophilus influenzae*, mtube – *Mycobacterium tuberculosis*, tpall – *Treponema pallidum*, rprow – *Rickettsia prowazekii*, mgeni – *Mycoplasma genitalium*, bburg – *Borrelia burgdorferi*, afulg – *Archaeoglobus fulgidus*, mjann – *Methanococcus jannaschii*, phori – *Pyrococcus horikoshii*, scere – *Saccharomyces cerevisiae*, spomb – *Schizosaccharomyces pombe*, pfalc – *Plasmodium falciparum*).

U nukleotiidi		↑: sagedus on KEG-ides tõusnud võrreldes kõigi geenidega. ↓: sagedus on KEG-ides langenud võrreldes kõigi geenidega.													
nukleotiidi		organism													
pos	ecoli	bsubt	hpylo	hinfl	mtube	tpall	rprow	mgeni	bburg	afulg	mjann	phori	scere	spomb	pfalc
4	0,108 ↑	0,306 ↓	0,301 ↓	0,513 ↑	0,001 ↓	0,052 ↓	0,851 ↓	0,340 ↓	0,703 ↑	0,477 ↓	0,396 ↓	0,808 ↑	0,305 ↑	0,127 ↓	1,000 ↓
5	0,099 ↓	0,122 ↓	0,011 ↓	0,037 ↓	0,003 ↓	0,084 ↓	0,007 ↓	0,350 ↓	0,036 ↓	0,059 ↓	0,016 ↓	0,068 ↓	0,128 ↓	0,168 ↓	0,482 ↓
6	0,037 ↑	0,545 ↑	0,070 ↓	0,036 ↑	0,865 ↑	0,414 ↑	0,219 ↓	0,621 ↓	0,361 ↓	1,000 ↓	0,372 ↓	1,000 ↓	1,0E-06 ↑	0,794 ↑	0,070 ↓
7	0,031 ↓	0,564 ↓	0,046 ↓	0,051 ↓	0,096 ↓	0,162 ↓	0,414 ↓	0,308 ↓	0,710 ↓	0,849 ↑	0,865 ↑	0,073 ↓	0,181 ↓	0,293 ↓	0,141 ↓
8	0,705 ↓	0,214 ↑	0,656 ↓	0,691 ↓	0,653 ↑	0,614 ↑	0,337 ↑	0,228 ↓	0,875 ↑	0,305 ↓	0,037 ↓	0,302 ↓	0,440 ↓	1,000 ↓	0,005 ↓
9	0,024 ↑	0,157 ↓	0,771 ↓	0,364 ↓	0,322 ↓	0,504 ↑	0,363 ↓	0,190 ↓	0,648 ↓	0,046 ↓	0,155 ↓	0,007 ↓	0,050 ↓	0,014 ↑	1,000 ↓
10	0,036 ↓	0,008 ↓	0,007 ↓	0,260 ↓	0,183 ↑	0,851 ↓	0,243 ↓	0,404 ↓	0,110 ↓	0,012 ↓	0,610 ↓	0,848 ↑	0,895 ↓	0,282 ↓	0,401 ↓
11	0,555 ↑	1,000 ↓	1,000 ↑	0,529 ↓	0,768 ↓	0,336 ↑	0,179 ↓	0,087 ↓	0,545 ↓	0,397 ↓	0,021 ↓	0,066 ↑	0,031 ↑	0,666 ↓	1,000 ↓
12	1,0E-03 ↑	0,648 ↑	0,771 ↓	0,258 ↑	0,733 ↓	0,871 ↑	1,000 ↓	0,022 ↓	0,444 ↓	0,410 ↑	0,142 ↓	1,000 ↑	0,007 ↑	0,346 ↓	0,073 ↓
13	1,000 ↓	0,488 ↓	0,072 ↓	0,042 ↓	0,201 ↓	0,364 ↓	0,394 ↓	0,446 ↑	0,064 ↓	0,851 ↑	0,276 ↓	0,123 ↓	0,001 ↓	0,019 ↓	0,238 ↓
14	0,407 ↑	0,559 ↑	1,000 ↑	0,315 ↑	0,314 ↓	0,267 ↑	0,880 ↓	0,409 ↓	0,230 ↓	0,022 ↓	0,098 ↓	0,008 ↓	0,524 ↓	0,550 ↑	0,632 ↑
15	0,180 ↓	0,368 ↑	0,079 ↓	0,134 ↑	0,861 ↑	0,109 ↑	0,378 ↑	1,000 ↑	1,000 ↑	0,754 ↑	0,663 ↓	0,634 ↑	0,335 ↑	0,417 ↑	0,100 ↓
16	0,020 ↓	0,865 ↓	0,875 ↓	0,031 ↓	0,457 ↓	0,380 ↓	0,732 ↑	0,170 ↓	0,519 ↓	0,012 ↓	0,869 ↓	1,000 ↓	1,000 ↓	0,097 ↓	0,183 ↓
17	0,906 ↓	0,887 ↑	0,393 ↓	0,377 ↓	0,239 ↑	0,426 ↓	0,881 ↑	0,011 ↑	0,097 ↓	0,574 ↓	0,073 ↓	0,114 ↓	0,710 ↓	0,772 ↓	0,638 ↓
18	0,144 ↑	0,167 ↑	0,668 ↑	0,706 ↑	0,363 ↓	0,530 ↑	0,762 ↓	0,628 ↑	0,368 ↓	0,528 ↑	1,000 ↑	0,007 ↑	6,0E-05 ↑	0,417 ↑	1,000 ↓
19	0,191 ↓	0,730 ↓	0,430 ↓	0,149 ↓	0,432 ↓	0,857 ↓	0,729 ↓	0,579 ↑	0,230 ↓	0,465 ↑	0,177 ↓	0,064 ↓	0,895 ↑	0,073 ↓	0,014 ↓
20	1,000 ↑	0,112 ↓	1,000 ↑	0,533 ↓	0,186 ↑	0,535 ↓	0,286 ↓	0,425 ↑	0,010 ↓	0,395 ↓	0,217 ↓	0,083 ↓	0,110 ↓	0,399 ↓	0,012 ↓
21	1,000 ↑	0,544 ↓	0,382 ↓	0,619 ↑	0,150 ↓	0,422 ↑	0,181 ↑	0,430 ↓	0,650 ↓	0,527 ↑	0,379 ↓	0,220 ↑	0,008 ↓	1,000 ↑	0,655 ↑
22	0,312 ↓	0,610 ↓	0,011 ↓	1,000 ↓	1,000 ↓	0,059 ↓	0,008 ↓	0,268 ↓	0,010 ↓	0,192 ↓	0,488 ↓	0,578 ↓	0,009 ↓	0,022 ↓	0,061 ↓
23	0,406 ↑	0,391 ↓	0,254 ↓	0,800 ↑	1,000 ↑	0,426 ↓	0,004 ↓	0,332 ↓	5,1E-05 ↓	0,312 ↑	0,091 ↓	0,151 ↓	0,017 ↓	0,392 ↓	0,644 ↓
24	0,805 ↓	0,175 ↑	0,660 ↓	0,618 ↓	0,352 ↑	0,520 ↑	0,883 ↑	0,637 ↓	1,000 ↑	1,000 ↓	0,309 ↓	0,121 ↑	0,116 ↑	0,181 ↑	0,881 ↑
25	0,002 ↓	1,000 ↓	0,349 ↓	0,037 ↓	0,079 ↓	0,084 ↓	0,163 ↓	0,575 ↓	0,141 ↓	0,194 ↓	0,328 ↑	0,274 ↓	0,414 ↓	0,059 ↓	0,021 ↓
26	0,811 ↓	0,772 ↓	0,476 ↓	1,000 ↓	0,455 ↑	0,521 ↓	0,294 ↓	0,748 ↑	1,000 ↓	0,243 ↓	0,671 ↓	0,002 ↓	0,207 ↑	0,322 ↓	0,286 ↓
27	0,383 ↓	0,881 ↓	0,105 ↓	0,311 ↓	1,000 ↓	0,520 ↑	0,459 ↑	1,000 ↓	0,293 ↓	0,408 ↓	0,475 ↓	0,438 ↓	0,469 ↓	0,059 ↓	0,369 ↑
28	0,100 ↓	0,215 ↓	0,154 ↓	0,017 ↓	0,841 ↓	0,459 ↓	0,489 ↓	0,560 ↓	0,045 ↓	0,144 ↓	0,408 ↓	0,143 ↓	0,137 ↓	1,1E-04 ↓	0,001 ↓
29	0,230 ↓	0,060 ↓	0,393 ↓	0,094 ↓	0,392 ↓	0,632 ↓	0,882 ↓	0,513 ↓	0,458 ↓	0,026 ↓	0,067 ↓	0,559 ↓	0,617 ↓	0,118 ↓	0,092 ↓
30	0,298 ↓	1,000 ↑	0,061 ↓	0,531 ↑	0,711 ↓	0,310 ↑	1,000 ↓	0,266 ↑	0,548 ↑	0,747 ↑	0,773 ↓	0,759 ↑	0,148 ↑	0,418 ↑	0,298 ↓

(jätkub)

C nukleotiid

↑: sagedus on KEG-ides tõusnud võrreldes kõigi geenidega. ↓: sagedus on KEG-ides langenud võrreldes kõigi geenidega.

nukleotiidi	organism														
pos	ecoli	bsubt	hpylo	hinf	mtube	tpall	rpro	mgeni	bburg	afulg	mjann	phori	scere	spomb	pfalc
4	0,240 ↓	0,272 ↑	0,064 ↓	0,185 ↓	0,632 ↑	0,006 ↓	0,051 ↓	1,000 ↓	0,079 ↓	0,125 ↓	0,465 ↑	1,000 ↓	0,506 ↓	1,000 ↓	0,506 ↓
5	2,2E-05 ↑	2,1E-05 ↑	3,4E-06 ↑	5,0E-06 ↑	2,6E-05 ↑	0,010 ↑	0,001 ↑	0,123 ↑	1,5E-04 ↑	3,0E-05 ↑	8,7E-07 ↑	6,3E-04 ↑	6,3E-06 ↑	0,001 ↑	6,3E-06 ↑
6	0,392 ↓	0,428 ↓	0,510 ↓	0,055 ↓	0,358 ↓	3,8E-05 ↓	0,792 ↓	0,416 ↓	0,375 ↑	0,470 ↓	1,000 ↓	0,163 ↑	0,313 ↓	0,711 ↓	0,313 ↓
7	0,673 ↓	0,582 ↓	0,134 ↑	0,033 ↑	0,879 ↑	1,000 ↓	0,530 ↓	0,820 ↓	0,762 ↑	1,000 ↓	0,179 ↓	0,040 ↑	0,876 ↓	0,073 ↑	0,876 ↑
8	1,000 ↑	0,256 ↓	1,000 ↓	0,201 ↑	0,211 ↓	0,050 ↓	0,120 ↑	0,379 ↑	0,781 ↑	0,698 ↑	0,052 ↑	0,838 ↓	0,363 ↓	0,882 ↑	0,363 ↑
9	0,004 ↓	1,000 ↑	1,000 ↑	0,306 ↓	0,503 ↑	0,151 ↓	1,000 ↓	1,000 ↓	0,355 ↓	0,089 ↓	0,380 ↑	0,080 ↑	0,025 ↑	0,477 ↓	0,025 ↓
10	0,030 ↓	0,214 ↓	0,829 ↓	0,038 ↓	0,119 ↓	0,861 ↓	0,824 ↓	0,673 ↑	0,791 ↓	0,331 ↓	0,285 ↑	0,672 ↓	0,143 ↓	0,395 ↑	0,143 ↑
11	0,680 ↓	0,854 ↑	0,854 ↓	0,751 ↓	0,784 ↑	0,730 ↓	0,387 ↓	0,525 ↑	0,814 ↑	0,089 ↓	0,119 ↑	0,673 ↓	0,694 ↑	0,767 ↓	0,694 ↓
12	0,413 ↓	0,075 ↓	0,373 ↓	0,363 ↓	0,105 ↑	0,722 ↓	0,793 ↓	0,817 ↑	0,772 ↑	0,331 ↓	0,321 ↑	0,228 ↓	0,380 ↑	0,004 ↑	0,380 ↓
13	0,014 ↓	0,214 ↓	0,168 ↓	0,434 ↓	0,001 ↓	0,090 ↓	1,000 ↓	0,820 ↓	0,245 ↓	0,331 ↓	0,314 ↓	0,309 ↑	0,561 ↑	0,742 ↓	0,561 ↓
14	0,040 ↓	0,576 ↑	1,000 ↓	0,649 ↓	0,395 ↑	1,000 ↓	0,842 ↑	0,824 ↑	0,599 ↑	0,052 ↓	0,823 ↑	0,439 ↑	0,235 ↑	0,088 ↑	0,235 ↑
15	0,688 ↑	0,170 ↑	1,000 ↓	0,868 ↑	0,790 ↑	0,090 ↓	0,793 ↓	0,101 ↓	0,355 ↓	0,039 ↓	0,800 ↑	0,705 ↓	0,549 ↓	0,016 ↑	0,549 ↓
16	0,606 ↑	0,711 ↑	0,701 ↑	0,236 ↓	0,884 ↑	1,000 ↑	1,000 ↓	0,823 ↑	0,201 ↑	1,000 ↓	1,000 ↓	0,319 ↓	0,367 ↓	0,863 ↓	0,367 ↑
17	0,500 ↓	0,026 ↓	0,848 ↑	1,000 ↓	0,124 ↓	0,098 ↓	0,670 ↑	0,636 ↓	0,644 ↓	0,269 ↑	0,043 ↑	0,438 ↑	0,092 ↑	0,355 ↑	0,092 ↑
18	0,344 ↓	0,258 ↓	0,373 ↓	0,617 ↓	0,691 ↑	1,000 ↓	0,570 ↓	0,809 ↑	0,573 ↓	0,420 ↓	0,637 ↑	0,717 ↓	0,652 ↑	0,004 ↑	0,652 ↓
19	0,237 ↑	0,098 ↑	0,690 ↑	1,000 ↓	0,569 ↓	0,730 ↓	0,672 ↑	0,432 ↑	0,795 ↑	0,854 ↓	0,087 ↑	0,697 ↑	0,300 ↓	0,226 ↑	0,300 ↑
20	1,000 ↓	0,582 ↓	0,453 ↑	0,341 ↑	0,492 ↓	0,217 ↓	1,000 ↓	0,832 ↓	0,486 ↑	0,263 ↑	0,051 ↑	0,564 ↑	0,138 ↑	0,878 ↓	0,138 ↓
21	0,795 ↑	0,597 ↑	0,585 ↑	0,402 ↓	0,235 ↑	0,476 ↓	0,617 ↓	0,811 ↑	0,429 ↑	1,000 ↓	0,038 ↓	0,717 ↓	3,8E-08 ↑	0,001 ↑	3,8E-08 ↑
22	1,000 ↓	1,000 ↓	0,850 ↑	1,000 ↓	0,160 ↓	0,050 ↓	0,198 ↑	1,000 ↓	0,603 ↑	0,354 ↑	1,000 ↓	0,336 ↑	0,471 ↑	0,042 ↑	0,471 ↑
23	1,000 ↓	0,109 ↓	0,274 ↓	0,458 ↓	0,891 ↓	1,000 ↓	0,059 ↓	1,000 ↓	0,833 ↑	0,855 ↓	1,000 ↓	0,849 ↑	0,509 ↓	1,000 ↓	0,509 ↑
24	0,440 ↑	0,859 ↑	0,353 ↓	0,185 ↓	0,290 ↑	0,232 ↑	0,603 ↑	0,168 ↑	0,032 ↑	0,201 ↓	0,083 ↓	0,423 ↓	0,058 ↓	0,862 ↓	0,058 ↓
25	0,431 ↓	0,597 ↓	0,227 ↓	0,014 ↓	0,567 ↑	0,080 ↓	0,662 ↑	0,262 ↑	0,327 ↑	0,848 ↓	0,084 ↓	0,020 ↑	0,771 ↓	0,141 ↓	0,771 ↑
26	0,790 ↓	0,475 ↑	0,090 ↓	0,885 ↓	0,064 ↓	0,862 ↑	0,687 ↑	0,520 ↓	0,835 ↓	1,000 ↓	0,843 ↓	0,137 ↑	0,189 ↓	0,761 ↑	0,189 ↑
27	0,432 ↑	0,078 ↑	0,003 ↑	0,869 ↑	0,115 ↓	0,586 ↓	0,171 ↓	0,514 ↑	0,247 ↓	0,108 ↓	0,635 ↑	0,853 ↓	0,112 ↑	1,9E-04 ↑	0,112 ↑
28	0,034 ↓	0,298 ↓	0,705 ↓	0,286 ↓	0,473 ↑	1,000 ↓	1,000 ↓	0,230 ↓	1,000 ↓	0,710 ↑	0,813 ↓	0,227 ↓	0,028 ↓	0,031 ↑	0,028 ↑
29	0,789 ↓	0,489 ↓	0,858 ↑	0,541 ↑	0,676 ↑	0,599 ↓	1,000 ↓	0,667 ↑	1,000 ↓	0,280 ↑	1,000 ↓	0,853 ↓	0,432 ↓	0,534 ↓	0,432 ↓
30	0,364 ↑	0,860 ↑	0,365 ↑	0,739 ↑	0,429 ↑	0,081 ↓	0,305 ↓	0,659 ↓	1,000 ↓	0,109 ↓	0,817 ↑	0,713 ↑	0,197 ↑	0,009 ↑	0,197 ↑

(jätkub)

A nukleotiid ↑: sagedus on KEG-ides tõusnud võrreldes kõigi geenidega. ↓: sagedus on KEG-ides langenud võrreldes kõigi geenidega.

nukleotiidi	organism														
pos	ecoli	bsubt	hpylo	hinfl	mtube	tpall	rpro	mgeni	bburg	afulg	mjann	phori	scere	spomb	pfalc
4	0,009 ↓	0,129 ↓	0,779 ↑	0,010 ↓	0,430 ↓	0,441 ↑	0,304 ↓	0,427 ↓	0,017 ↓	0,267 ↓	0,020 ↓	0,575 ↓	2,3E-06 ↓	0,044 ↓	0,002 ↓
5	0,057 ↓	0,003 ↓	0,120 ↓	0,008 ↓	0,478 ↓	0,751 ↑	0,006 ↓	0,212 ↓	0,006 ↓	0,197 ↓	0,016 ↓	0,302 ↓	4,7E-04 ↓	3,8E-04 ↓	4,7E-07 ↓
6	0,110 ↑	0,199 ↑	0,008 ↑	0,540 ↑	0,100 ↑	0,151 ↑	0,304 ↑	0,350 ↑	0,227 ↑	0,351 ↓	0,013 ↑	0,879 ↓	0,004 ↓	0,099 ↑	0,002 ↑
7	0,422 ↑	0,782 ↓	0,069 ↑	0,625 ↓	0,004 ↑	0,208 ↑	0,883 ↑	0,280 ↑	0,656 ↑	0,385 ↑	0,786 ↑	0,888 ↑	0,478 ↑	1,000 ↓	0,240 ↑
8	0,643 ↑	0,783 ↑	0,783 ↑	0,046 ↓	0,104 ↑	0,186 ↑	0,077 ↓	0,639 ↓	0,457 ↓	0,040 ↑	0,671 ↓	0,026 ↑	0,906 ↑	0,782 ↓	0,381 ↑
9	0,711 ↑	0,002 ↑	0,569 ↑	0,019 ↑	0,721 ↑	0,853 ↑	0,379 ↑	0,115 ↑	0,765 ↑	0,210 ↑	0,326 ↑	0,761 ↑	0,185 ↑	0,048 ↓	0,364 ↑
10	7,8E-05 ↑	0,258 ↑	0,016 ↑	0,062 ↑	0,070 ↑	0,137 ↑	0,106 ↑	0,640 ↑	0,139 ↑	0,042 ↑	0,684 ↓	0,158 ↑	0,344 ↓	0,268 ↓	0,553 ↑
11	0,459 ↑	0,669 ↓	0,774 ↑	0,523 ↑	0,205 ↑	1,000 ↓	0,051 ↑	0,117 ↑	0,224 ↑	0,092 ↑	0,885 ↓	0,209 ↓	0,018 ↓	0,494 ↑	0,306 ↑
12	0,541 ↑	0,015 ↑	0,305 ↑	0,536 ↑	0,464 ↑	0,584 ↑	0,380 ↑	0,208 ↑	1,000 ↑	0,145 ↓	0,396 ↑	0,210 ↓	0,076 ↓	0,492 ↓	0,023 ↑
13	0,007 ↑	0,477 ↑	0,047 ↑	0,045 ↑	1,4E-04 ↑	0,001 ↑	0,558 ↑	0,755 ↓	0,002 ↑	0,464 ↑	0,497 ↑	0,472 ↓	0,293 ↑	0,013 ↑	0,106 ↑
14	0,013 ↑	0,476 ↓	0,564 ↑	0,606 ↓	0,033 ↑	0,862 ↑	0,652 ↑	0,752 ↑	0,446 ↑	0,174 ↑	0,400 ↑	0,174 ↑	0,243 ↑	0,596 ↓	0,240 ↓
15	0,055 ↑	0,775 ↓	0,466 ↑	0,703 ↓	0,199 ↓	1,000 ↓	0,766 ↑	0,211 ↑	0,880 ↑	1,000 ↑	0,488 ↑	0,646 ↑	0,813 ↑	0,007 ↓	0,175 ↑
16	0,548 ↓	0,085 ↓	0,009 ↓	0,701 ↓	0,062 ↑	0,398 ↑	0,769 ↓	1,000 ↑	0,653 ↓	0,661 ↓	0,276 ↑	0,080 ↓	0,724 ↓	0,891 ↓	0,659 ↑
17	1,000 ↓	0,242 ↑	0,382 ↑	0,435 ↑	0,019 ↑	0,020 ↑	0,292 ↓	0,530 ↓	0,044 ↑	0,878 ↓	0,670 ↑	0,219 ↑	1,000 ↓	0,580 ↓	0,025 ↓
18	1,000 ↑	0,666 ↓	0,765 ↓	0,613 ↑	0,177 ↓	0,356 ↑	0,303 ↑	0,874 ↓	0,170 ↑	0,737 ↓	0,784 ↑	0,267 ↓	0,003 ↓	0,671 ↓	0,066 ↑
19	0,902 ↓	0,666 ↓	0,669 ↑	0,699 ↑	0,024 ↑	0,169 ↑	0,882 ↓	0,874 ↓	0,657 ↑	0,369 ↓	0,412 ↓	0,772 ↑	0,639 ↓	0,781 ↑	0,769 ↓
20	0,614 ↑	0,104 ↑	0,772 ↑	0,364 ↓	0,161 ↑	0,195 ↑	0,881 ↑	0,869 ↑	0,213 ↑	1,000 ↓	1,000 ↑	0,164 ↑	0,074 ↑	0,166 ↑	0,055 ↑
21	0,353 ↓	0,242 ↑	0,768 ↑	0,702 ↑	1,000 ↓	0,448 ↑	0,769 ↓	0,411 ↑	0,762 ↓	0,869 ↓	0,074 ↑	0,650 ↓	0,464 ↓	0,003 ↓	0,763 ↑
22	0,805 ↓	0,384 ↑	0,305 ↑	0,794 ↓	0,250 ↑	0,025 ↑	0,137 ↓	0,588 ↑	0,551 ↑	0,052 ↑	0,338 ↑	0,663 ↑	0,724 ↓	0,676 ↓	0,771 ↓
23	0,899 ↑	0,183 ↑	0,139 ↑	0,794 ↑	0,083 ↑	0,055 ↑	0,532 ↑	0,617 ↑	0,002 ↑	0,442 ↓	0,671 ↑	0,087 ↑	0,406 ↑	0,680 ↓	0,378 ↓
24	1,000 ↓	0,459 ↓	0,378 ↑	0,157 ↑	0,841 ↓	0,244 ↓	0,882 ↑	0,424 ↓	0,362 ↓	0,317 ↑	0,028 ↑	0,218 ↓	0,118 ↑	0,666 ↓	0,453 ↑
25	0,262 ↑	0,657 ↓	0,242 ↑	0,118 ↑	0,046 ↑	0,048 ↑	0,555 ↓	0,532 ↑	0,880 ↑	0,079 ↑	0,051 ↓	1,000 ↑	0,722 ↑	0,050 ↑	2,0E-04 ↑
26	0,797 ↑	1,000 ↓	1,000 ↓	0,894 ↓	0,231 ↑	0,861 ↓	0,645 ↓	0,609 ↓	0,527 ↓	0,132 ↑	0,888 ↓	0,878 ↑	0,906 ↑	0,264 ↑	1,000 ↓
27	0,587 ↑	0,372 ↑	0,878 ↓	0,307 ↑	0,545 ↓	0,700 ↓	0,765 ↓	0,739 ↓	0,449 ↓	0,620 ↓	0,098 ↑	1,000 ↓	0,460 ↑	0,029 ↓	0,881 ↓
28	0,022 ↑	0,233 ↑	0,039 ↑	0,041 ↓	0,524 ↑	0,167 ↑	0,657 ↓	0,622 ↑	1,000 ↓	0,224 ↑	0,491 ↑	0,462 ↑	0,031 ↑	1,000 ↓	0,458 ↑
29	0,157 ↑	0,177 ↑	0,762 ↑	0,427 ↑	0,147 ↑	1,000 ↓	1,000 ↑	1,000 ↑	0,424 ↓	0,366 ↑	0,478 ↑	0,879 ↓	0,630 ↓	0,780 ↑	1,000 ↓
30	0,489 ↑	0,033 ↑	1,000 ↑	0,406 ↑	0,404 ↓	0,852 ↑	0,136 ↑	0,630 ↓	0,880 ↓	0,866 ↑	0,681 ↑	0,649 ↑	0,038 ↓	0,059 ↓	0,879 ↓

(jätkub)

G nukleotiid

↑: sagedus on KEG-ides tõusnud võrreldes kõigi geenidega. ↓: sagedus on KEG-ides langenud võrreldes kõigi geenidega.

nukleotiidi	organism														
pos	ecoli	bsubt	hpylo	hinfl	mtube	tpall	rpro	mgeni	bburg	afulg	mjann	phori	scere	spomb	pfalc
4	0,005 ↑	0,049 ↑	0,024 ↑	0,001 ↑	0,008 ↑	0,005 ↑	2,2E-04 ↑	0,064 ↑	4,6E-04 ↑	0,004 ↑	0,006 ↑	0,541 ↑	1,7E-04 ↑	8,1E-04 ↑	0,001 ↑
5	0,366 ↓	0,553 ↑	1,000 ↓	1,000 ↓	0,118 ↓	0,255 ↓	0,003 ↑	0,113 ↑	0,014 ↑	0,287 ↓	0,134 ↑	0,734 ↑	0,875 ↓	0,167 ↑	0,224 ↑
6	7,8E-05 ↓	0,107 ↓	0,673 ↓	0,013 ↓	0,673 ↓	0,240 ↑	0,285 ↑	0,588 ↑	0,386 ↓	0,106 ↑	0,074 ↓	0,553 ↓	0,068 ↓	0,061 ↓	0,674 ↓
7	0,142 ↑	0,064 ↑	0,256 ↓	0,449 ↑	0,111 ↓	1,000 ↓	0,157 ↑	0,836 ↑	1,000 ↓	0,471 ↓	0,435 ↑	0,879 ↓	0,375 ↑	0,886 ↓	0,863 ↓
8	1,000 ↓	0,676 ↓	0,497 ↑	0,836 ↑	0,628 ↓	1,000 ↓	1,000 ↓	0,155 ↑	0,382 ↑	0,144 ↓	0,096 ↑	0,373 ↓	0,030 ↑	0,698 ↑	0,380 ↑
9	0,893 ↓	0,033 ↓	0,847 ↓	0,515 ↓	1,000 ↑	0,640 ↑	0,637 ↑	0,806 ↓	0,393 ↑	0,031 ↑	1,000 ↓	0,544 ↑	0,100 ↓	1,000 ↓	1,000 ↓
10	0,781 ↓	0,031 ↑	0,866 ↑	0,342 ↑	0,274 ↓	0,521 ↓	0,834 ↓	1,000 ↓	0,857 ↑	0,558 ↑	0,530 ↑	0,209 ↓	0,008 ↑	0,106 ↑	1,000 ↓
11	0,267 ↓	0,405 ↑	1,000 ↓	0,425 ↑	0,337 ↓	0,722 ↓	0,568 ↑	0,608 ↓	0,385 ↓	0,847 ↓	0,034 ↑	0,838 ↓	1,000 ↓	0,688 ↑	1,000 ↓
12	1,7E-05 ↓	0,026 ↓	0,839 ↑	0,084 ↓	0,055 ↓	1,000 ↓	0,517 ↓	0,141 ↑	0,395 ↑	0,133 ↑	1,000 ↓	0,039 ↑	0,032 ↓	0,586 ↓	0,814 ↓
13	0,686 ↓	0,330 ↑	0,512 ↑	0,648 ↑	0,689 ↑	0,513 ↓	0,695 ↑	0,692 ↑	0,437 ↓	1,000 ↓	0,454 ↑	0,229 ↑	0,298 ↑	1,000 ↓	0,852 ↑
14	0,049 ↓	1,000 ↓	0,510 ↓	1,000 ↓	0,022 ↓	0,164 ↓	0,763 ↓	0,616 ↑	0,635 ↑	0,004 ↑	0,273 ↑	0,335 ↑	0,021 ↓	0,057 ↓	0,653 ↓
15	0,392 ↓	0,108 ↓	0,111 ↑	0,027 ↓	0,493 ↑	1,000 ↓	0,169 ↓	1,000 ↓	0,658 ↑	0,132 ↑	0,738 ↓	0,755 ↓	0,460 ↓	1,000 ↓	0,488 ↑
16	0,016 ↑	0,083 ↑	0,005 ↑	0,136 ↑	0,231 ↓	1,000 ↓	0,688 ↑	0,236 ↑	0,577 ↑	0,012 ↑	0,536 ↓	0,228 ↑	0,153 ↑	0,023 ↑	0,583 ↑
17	0,205 ↑	0,535 ↑	1,000 ↓	0,699 ↑	0,052 ↓	0,736 ↑	0,302 ↑	1,000 ↓	0,641 ↑	1,000 ↑	1,000 ↓	1,000 ↓	0,270 ↓	0,729 ↑	0,827 ↑
18	0,586 ↓	0,729 ↑	0,340 ↑	0,679 ↓	0,139 ↑	0,324 ↓	0,791 ↓	0,808 ↓	1,000 ↑	0,442 ↑	0,581 ↓	0,267 ↓	0,051 ↓	0,322 ↓	0,269 ↓
19	0,800 ↑	0,876 ↓	0,748 ↑	0,251 ↑	0,510 ↓	0,633 ↓	0,583 ↑	0,307 ↓	0,581 ↑	0,561 ↑	0,293 ↑	0,441 ↑	0,183 ↑	0,655 ↑	0,031 ↑
20	0,647 ↓	0,310 ↑	0,198 ↓	0,173 ↑	0,031 ↓	0,382 ↑	0,081 ↑	0,340 ↓	0,147 ↑	0,852 ↑	1,000 ↓	1,000 ↑	0,026 ↓	0,016 ↓	0,066 ↑
21	0,408 ↑	0,377 ↓	0,849 ↑	1,000 ↓	1,000 ↑	0,627 ↓	0,571 ↓	1,000 ↓	0,521 ↑	0,875 ↓	0,711 ↑	1,000 ↓	0,058 ↓	0,849 ↓	0,269 ↓
22	0,129 ↑	0,878 ↓	0,267 ↑	0,481 ↑	0,693 ↑	0,274 ↑	0,002 ↑	0,857 ↓	0,106 ↑	0,192 ↓	1,000 ↓	0,548 ↓	0,155 ↑	0,375 ↑	0,360 ↑
23	0,298 ↓	0,151 ↑	0,401 ↑	0,706 ↑	0,199 ↓	0,476 ↓	0,358 ↓	0,390 ↑	0,320 ↑	0,719 ↓	0,078 ↑	0,847 ↓	0,015 ↑	0,071 ↑	0,654 ↑
24	0,785 ↓	0,610 ↓	0,454 ↑	0,700 ↑	0,140 ↓	0,627 ↓	0,420 ↓	0,387 ↑	1,000 ↓	0,645 ↓	0,854 ↓	1,000 ↓	0,058 ↓	0,734 ↓	0,479 ↓
25	0,052 ↑	0,224 ↑	0,432 ↑	0,018 ↑	0,361 ↓	0,222 ↑	0,045 ↑	0,446 ↓	0,472 ↑	0,884 ↓	0,017 ↑	0,366 ↓	0,441 ↑	0,306 ↑	0,716 ↓
26	0,557 ↑	1,000 ↑	0,005 ↑	0,479 ↑	1,000 ↓	0,295 ↑	0,442 ↑	0,822 ↑	0,169 ↑	1,000 ↓	0,340 ↑	0,053 ↑	1,000 ↓	0,857 ↓	0,241 ↑
27	0,793 ↓	0,016 ↓	0,468 ↓	1,000 ↓	0,347 ↓	0,639 ↑	0,109 ↑	0,800 ↑	0,162 ↑	0,003 ↑	0,144 ↓	0,179 ↑	0,153 ↓	0,311 ↑	0,115 ↓
28	0,226 ↑	0,370 ↑	0,875 ↓	0,002 ↑	0,512 ↓	0,753 ↓	1,000 ↑	0,227 ↑	0,023 ↑	0,885 ↓	0,764 ↑	0,182 ↑	0,306 ↑	0,139 ↑	0,026 ↑
29	0,662 ↑	0,076 ↑	0,432 ↑	0,509 ↑	0,523 ↓	0,732 ↑	0,169 ↓	0,645 ↑	0,024 ↑	0,594 ↑	0,128 ↑	0,138 ↑	0,013 ↑	0,013 ↑	0,010 ↑
30	0,248 ↓	0,007 ↓	0,145 ↑	0,013 ↓	0,790 ↑	0,635 ↑	0,868 ↓	1,000 ↓	0,824 ↓	0,286 ↑	1,000 ↓	0,309 ↓	0,666 ↓	0,095 ↓	0,530 ↑

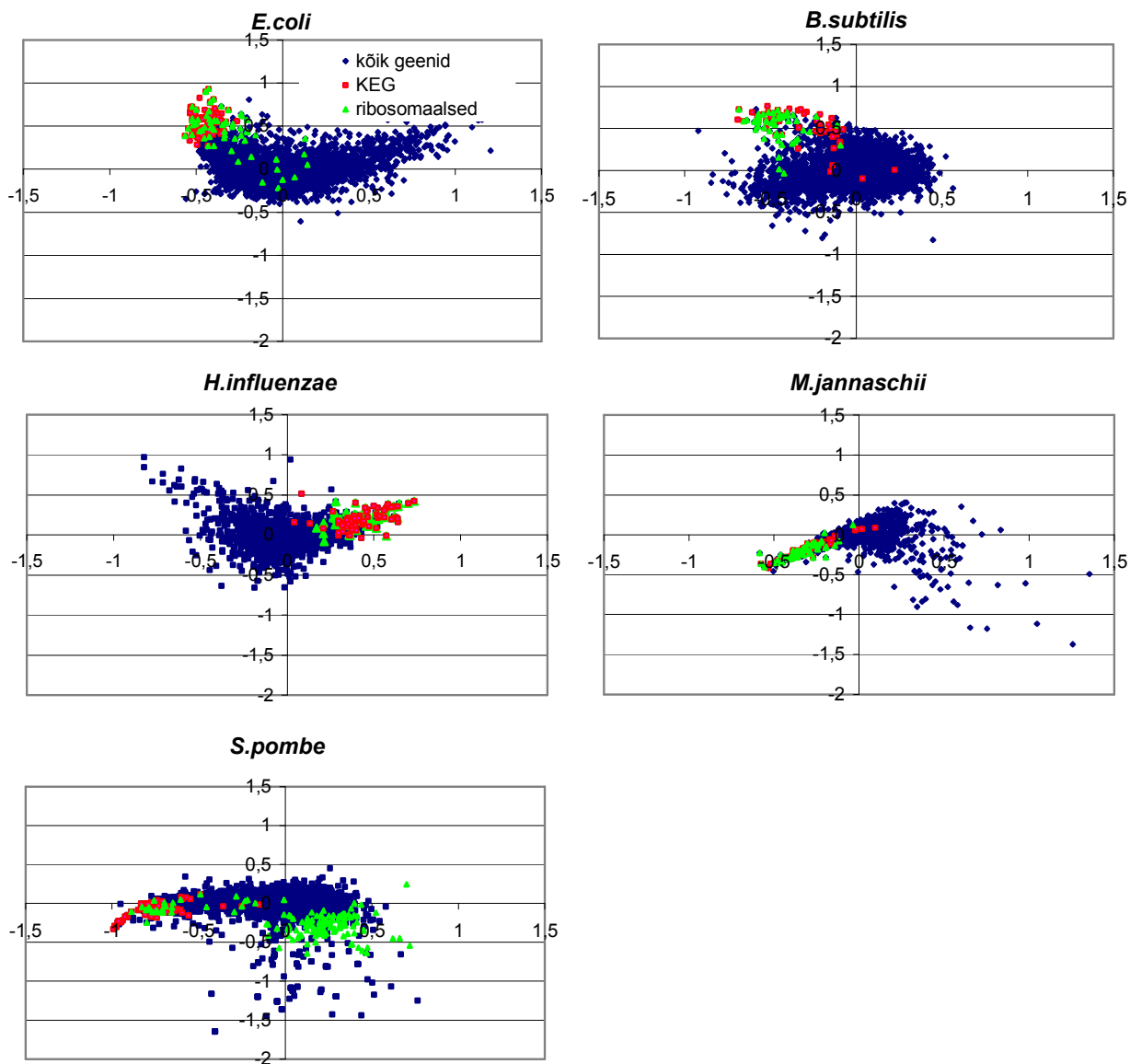
Lisa IV. Juhuslikult valitud organismide *E.coli* KEG-idele vastavad ortoloogid.

<i>E.coli</i> (80)	<i>H.marismortui</i> (33)	<i>M.avium</i> <i>paratuberculosis</i> (63)	<i>R.conorii</i> (53)	<i>S.elongatus</i> (59)	<i>E.coli</i> (80)	<i>H.marismortui</i> (33)	<i>M.avium</i> <i>paratuberculosis</i> (63)	<i>R.conorii</i> (53)	<i>S.elongatus</i> (59)
lpp					rplK	+	+	+	+
rplL		+	+	+	rplO		+	+	+
eno	+	+		+	rplD		+	+	+
gapA	+	+		+	rpoC	+	+	+	+
ompC					rpsO		+		+
tufA	+				cspC	+		+	
cspA	+	+			efp		+	+	+
ahpC	+	+	+	+	slyD	+			
tufB		+	+	+	pal			+	
mopA(groL)	+	+	+	+	trpL				
ompA					tktA		+		+
rpsI		+	+	+	rplT		+	+	+
rpsA		+		+	rpsF		+	+	+
pflB					aceE		+		
rpsB		+	+	+	atpA		+	+	+
fba				+	ppiB	+	+		+
tsf		+	+	+	acpP		+	+	+
rplA	+	+	+	+	pta	+	+	+	
rplY		+	+		rpsR		+	+	+
fusA	+	+	+	+	glyA	+	+	+	+
rpmA		+	+	+	hupA		+	+	+
tpiA		+		+	rplM	+	+	+	+
rpmE		+		+	rpsT		+	+	+
ompX					ompF				
tig		+	+	+	pfkA		+		+
pgk	+	+		+	rpsL	+	+	+	+
rpsC	+	+		+	ackA		+	+	+
rpmG		+	+		rpsP		+	+	+
rpsU				+	ppa	+	+	+	+
rpmH		+	+	+	deoD	+			
rplI		+	+	+	lpdA	+	+	+	+
dnaK	+	+	+	+	adhE	+	+	+	
sodA	+	+	+	+	rplW	+	+	+	+
rpmI		+		+	atpD	+	+	+	+
yaiU					adk	+	+	+	+
rplB	+	+		+	ybaB		+	+	+
rplC	+	+	+	+	bipA		+	+	+
yfiD					uxuA				
rpsQ	+	+	+	+	ndk	+	+	+	+
pykF	+	+		+	yebC		+	+	+

Lisa V. Juhuslikult valitud organismide *S.cerevisiae* KEG-idele vastavad ortoloogid.

<i>S.cerevisiae</i> (80)	<i>C.neoformans</i> (64)	<i>H.marismortui</i> (34)	<i>S.cerevisiae</i> (80)	<i>C.neoformans</i> (64)	<i>H.marismortui</i> (34)
TDH3			RPS9B	+	+
RPL39	+	+	RPS17A	+	+
TDH2			RPL17A	+	+
PDC1	+	+	RPS14A	+	+
CDC19	++		RPP2A	+	+
ENO2			EFT1	+	
TEF2	+		RPL4A		
RPL43A	+	+	SSA2	++	
ENO1	+		ILV5	+	+
TEF1			RPS10B		
CCW12			RPS2	+	
FBA1	+		EFT2		
RPL30	+		RPS3	+	
RPL22A	+		RPL36B	+	
TDH1	+		RPL35A	+	
RPL8B			RPL42B		
RPS6B	+	+	RPP0	+	+
RPS12	+		RPS19A	+	+
RPL8A	+	+	RPS23B	+	+
RPS10A	+		RPL15A	+	+
RPS5	+	+	RPS26A	+	
RPS6A			RPL34B	+	
RPL1B	+		RPL11A	+	+
RPL3	+		YEF3	+	+
RPL5	+	+	RPL9B	+	+
RPS20	+	+	ASC1	+	
RPL10	+	+	RPS13	+	+
RPL1A			RPS18A	+	+
SSB1			RPL38		
RPL32	+	+	RPL2A	+	
TPI1	+		SSB2	+	
RPS24A	+	+	RPL9A		
PGK1	++		RPS1B	+	+
HYP2	+	+	RPL42A		
RPL4B	+	+	RPS15	+	+
RPS22A	+	+	RPL12B	+	
RPL18A	+		CWP2		
GPM1		+	RPS16B	+	+
RPS31	+		RPL2B		
ADH1	+	+	EFB1	+	

Lisa VI. Näiteid erinevate organismide sünonüümse koodonkasutuse korrespondentsanalüüsi graafikutest.



Lisa VII. Lisaks analüüsitud kõrge ekspressioonitasemega geenid.

<i>B.subtilis</i>			<i>H.influenzae</i>				<i>M.jannaschii</i>			<i>S.pombe</i>		
abrB	rpmD	ykwD	HI0084	HI0585	HI0951	HI1609	MJ0032	MJ0476	MJ1168	SPAC1002.13c	SPAC4F8.07c	SPBC530.10c
aprE	rpmGA	ymzB	HI0110	HI0591	HI0968	HI1633	MJ0040	MJ0477	MJ1169	SPAC1002.17c	SPAC521.05	SPBC56F2.02
atpE	rpmH	yneF	HI0119	HI0620	HI0971	HI1634	MJ0049	MJ0498	MJ1171	SPAC1002.19	SPAC57A7.04c	SPBC660.16
atpF	rpmI	ynzC	HI0131	HI0640	HI0972	HI1659	MJ0081	MJ0546	MJ1186	SPAC1006.07	SPAC664.04c	SPBC685.06
clpX	rpoA	yocJ	HI0142	HI0646	HI0980	HI1693	MJ0083	MJ0595	MJ1192.1	SPAC1071.08	SPAC664.05	SPBC776.01
comX	rpoB	yodC	HI0146	HI0682	HI0989	HI1702	MJ0097	MJ0608	MJ1199	SPAC1071.10c	SPAC6G10.11c	SPBC800.04c
cspD	rpsD	yonB	HI0155	HI0686	HI1034	HI1711	MJ0098	MJ0648	MJ1201	SPAC11D3.01c	SPAC890.08	SPBC839.05c
dltC	rpsE	yonC	HI0158	HI0693	HI1050	HI1734	MJ0112	MJ0655	MJ1202	SPAC11G7.04	SPAC8C9.08	SPBC839.13c
fbaA	rpsF	yonK	HI0170	HI0736	HI1053		MJ0113	MJ0657	MJ1203	SPAC13G6.02c	SPAC9.09	SPBC8D2.03c
flgE	rpsG	yopL	HI0171	HI0738	HI1061		MJ0168	MJ0659	MJ1258	SPAC144.11	SPAC926.04c	SPBC8D2.04
flil	rpsH	yoqO	HI0201	HI0740	HI1062		MJ0199	MJ0673	MJ1271	SPAC1687.06c	SPAC959.08	SPBC8D2.18c
frr	rpsJ	yosA	HI0221	HI0746	HI1069		MJ0212	MJ0691	MJ1273	SPAC1783.08c	SPACUNK4.17	SPBP8B7.03c
fusA	rpsK	yppF	HI0221.1	HI0757	HI1075		MJ0216	MJ0707	MJ1277	SPAC1805.11c	SPAPB24D3.07c	SPBP8B7.06
glnA	rpsM	yqbN	HI0222	HI0776	HI1076		MJ0217	MJ0728	MJ1325	SPAC1805.13	SPAPB8E5.06c	SPCC1020.06c
groES	rpsR	yqdA	HI0229	HI0782	HI1124		MJ0220	MJ0746	MJ1346	SPAC1834.03c	SPBC106.18	SPCC1223.09
guaB	rpsS	yqeY	HI0241	HI0784	HI1125		MJ0221	MJ0784	MJ1391	SPAC1834.04	SPBC1105.12	SPCC1259.01c
hag	rpsT	yrbF	HI0246	HI0788	HI1172		MJ0222	MJ0800	MJ1419	SPAC18G6.14c	SPBC14F5.05c	SPCC1281.06c
infA	rpsU	yrcH	HI0250	HI0789	HI1193		MJ0226	MJ0822	MJ1455	SPAC19G12.06c	SPBC1604.05	SPCC1322.10
lysS	secE	yrcE	HI0315	HI0790	HI1204		MJ0242	MJ0825	MJ1509	SPAC19G12.08	SPBC1685.09	SPCC1322.11
lytA	sodA	yuaB	HI0317	HI0791	HI1210		MJ0268	MJ0842	MJ1511	SPAC1A6.04c	SPBC1685.13	SPCC1393.03
murAA	spolIQ	yuaJ	HI0323	HI0792	HI1211		MJ0269	MJ0845	MJ1534	SPAC1F12.02c	SPBC16H5.02	SPCC1672.02c
pdhA	sspA	yugI	HI0325	HI0793	HI1232		MJ0280	MJ0846	MJ1573	SPAC1F7.13c	SPBC17G9.10	SPCC1682.14
pdhB	sspB	yuiA	HI0348	HI0794	HI1252		MJ0285	MJ0854	MJ1575	SPAC1F8.07c	SPBC18E5.06	SPCC1739.13
pdhC	sspC	yusA	HI0362	HI0795	HI1264		MJ0299	MJ0863	MJ1601	SPAC23A1.11	SPBC18H10.12c	SPCC191.07
ppiB	sspJ	yvaB	HI0370	HI0796	HI1284		MJ0307	MJ0868	MJ1656	SPAC23C11.05	SPBC19F8.08	SPCC24B10.09
pstS	tatAY	yvcE	HI0371	HI0800	HI1302		MJ0308	MJ0926	MJ1670	SPAC23H4.06	SPBC19G7.03c	SPCC31H12.04c
ptsH	trxA	yzvB	HI0377	HI0801	HI1318		MJ0318	MJ0932		SPAC24C9.12c	SPBC1E8.05	SPCC330.06c
rplE	yaaD	ywkA	HI0411	HI0803	HI1333		MJ0327	MJ0959		SPAC24H6.07	SPBC215.05	SPCC330.14c
rplF	ybfE	yydF	HI0445	HI0808	HI1344		MJ0362	MJ0960		SPAC25B8.12c	SPBC21C3.13	SPCC364.03
rplJ	yddQ		HI0483	HI0809	HI1354		MJ0386	MJ0990		SPAC26A3.04	SPBC28F2.03	SPCC417.08
rplN	ydfP		HI0484	HI0824	HI1363		MJ0394	MJ1035		SPAC26H5.10c	SPBC29A10.08	SPCC553.10
rplP	yeth		HI0494	HI0834	HI1367		MJ0396	MJ1045		SPAC27E2.11c	SPBC29B5.03c	SPCC576.03c
rplQ	yfhD		HI0495	HI0835	HI1390		MJ0401	MJ1090		SPAC29A4.02c	SPBC2F12.04	SPCC576.11
rplR	yhbJ		HI0515	HI0843	HI1391		MJ0402	MJ1091		SPAC31G5.03	SPBC2G5.05	SPCC613.06
rplS	yhcN		HI0519	HI0847	HI1396		MJ0404	MJ1099		SPAC31G5.17c	SPBC32H8.12c	SPCC622.08c
rplU	yhfD		HI0534	HI0880	HI1398		MJ0458.1	MJ1118		SPAC343.12	SPBC336.10c	SPCC622.09
rplV	yjbD		HI0542	HI0916	HI1434.1		MJ0459	MJ1142		SPAC3A12.10	SPBC354.12	SPCC622.12c
rplW	yjcS		HI0553	HI0924	HI1522.1		MJ0466	MJ1148		SPAC3H5.05c	SPBC365.03c	SPCC622.18
rplX	yjdB		HI0557	HI0927	HI1532		MJ0467	MJ1153		SPAC3H5.07	SPBC3D6.02	SPCC736.15
rpmB	yjID		HI0572	HI0928	HI1545		MJ0468	MJ1156		SPAC3H5.12c	SPBC3D6.15	
rpmC	ykuU		HI0580	HI0937	HI1573		MJ0473	MJ1167		SPAC4F10.20	SPBC428.05c	
Kokku: 111			Kokku: 131				Kokku: 108			Kokku: 121		